The Use of Molecular Modelling to Study Enzymic Action

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Chemistry

at the

University of Canterbury

by

Wanting Jiao

August 2011

Table of Contents

Acknowledgement

I would like to thank Associate Professor Emily J. Parker and Emeritus Professor James M. Coxon for their help and guidance in this PhD project. I would also like to thank Dr Geoffrey Jameson from Massey University, for the helpful discussions on crystal structure refinement and data processing; thank Dr Quentin McDonald, who helped me get started in the world of molecular modelling, and gave useful advice on modelling methodologies.

I would also like to acknowledge Dr Richard Hutton and Miss Nicola Blackmore for conducting the experiments needed to verify modelling results. Thank all members of the Parker group for their friendly support.

Finally I would like to thank the tertiary education commission for providing the top achiever doctoral scholarship and the travel fund that enabled me to attend to several academic conferences.

Abstract

Molecular modelling has become widely used in chemistry and biology. The aim of this project is to use a range of molecular modelling techniques to study enzymic actions. This thesis consists of two parts.

Part A of this thesis describes computational studies conducted for the calpain-calpastatin system. Calpain is a cysteine protease. Over-expression of calpain is associated with many diseases. Calpastatin is the naturally occurring specific regulator of calpain activity. In this part of the thesis, the dynamic conformational preferences of region B of the inhibitory domain in calpastatin were examined in detail by using molecular dynamics simulations and stochastic dynamic simulations with Monte Carlo sampling.

Part B of the thesis explores the structure and function of the enzyme 3-dexoy-D-*arabino*-heptulosonate 7-phosphate synthase from *Mycobacterium tuberculosis* (*Mtu*DAH7PS). *Mtu*DAH7PS catalyses the first reaction of the shikimate pathway and is a target for the development of anti-tuberculosis drugs. *Mtu*DAH7PS is found to be synergistically inhibited by combinations of aromatic amino acids (Trp+Phe or Trp+Tyr), but not by any single aromatic amino acids. In this part of the thesis, this unique mechanism of allosteric regulation in *Mtu*DAH7PS was investigated by using a range molecular modelling techniques. Firstly protein crystal structure refinements were conducted and those crystal structures of *Mtu*DAH7PS in complex with various ligand molecules are described in Chapter 4. Secondly, the reaction mechanism and roles of active site residues were investigated in Chapter 5, through docking calculations (both rigid docking and induced fit docking) of a series of designed active site inhibitors. Finally, Chapter 6 discusses the molecular basis of the communication mechanism of allosteric regulation in *Mtu*DAH7PS.

List of Abbreviations

CM:	chorismate mutase
CP1B:	calpastatin inhibitory domain region B 20-mer peptide
DAH7P:	3-deoxy-D-*arabino*-heptulosonate 7-phosphate
DAH7PS:	3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase
E4P:	erythrose 4-phosphate
ITC:	isothermal calorimetry
MCSD:	stochastic dynamics with Monte Carlo sampling
MD:	molecular dynamics
MM:	molecular mechanics
NMR:	nuclear magnetic resonance
PBC:	periodic boundary condition
PDB:	protein data bank
PEP:	phosphoenolpyruvate
PKA:	protein kinase A
PKC:	protein kinase C
PME:	particle-mesh Ewald
QM:	quantum mechanics
RMSD:	root mean square distance
RMSF:	root mean square fluctuation
SCA:	statistical coupling analysis
TB:	tuberculosis
WT:	wild type

Amino acid one-letter and three-letter codes

A, Ala	alaine	E, Glu	glutamate	L, Leu	leucine
S, Ser	serine	R, Arg	arginine	Q, Gln	glutamine
K, Lys	lysine	T, Thr	threonine	N, Asn	asparagine
G, Gly	glycine	M, Met	methionine	W, Trp	tryptophan
D, Asp	aspartate	H, His	histidine	F, Phe	phenylalaine
T, Tyr	tyrosine	C, Cys	cystein	I, Ile	isoleucine
P, Pro	proline	V, Val	valine		

List of Figures

List of Tables

List of publications arising from the studies described in this PhD project (at date of examination)

1. Jiao, W., McDonald, Q., Coxon, J. M., and Parker, E. J. (2010) Molecular modeling studies of peptide inhibitors highlight the importance of conformational prearrangement for inhibition of calpain, *Biochemistry 49*, 5533-5539.

2. Webby, C. J., Jiao, W., Hutton, R. D., Blackmore, N. J., Baker, H. M., Baker, E. N., Jameson, G. B., and Parker, E. J. (2010) Synergistic allostery, a sophisticated regulatory network for the control of aromatic amino acid biosynthesis in *Mycobacterium tuberculosis*, *J. Biol. Chem. 285*, 30567-30576.

3. Reichau, S., Jiao, W., Walker, S. R., Hutton, R. D., Baker, E. N., and Parker, E. J. (2011) Potent inhibitors of a Shikimate pathway enzyme from *Mycobacterium tuberculosis*: combining mechanism- and modeling-based design, *J. Biol. Chem. 286*, 16197-16207.

4. Walker, S. R., Jiao, W., and Parker, E. J. (2011) Synthesis and evaluation of dual site inhibitors of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase, *Bioorg. Med. Chem. Lett. 21*, 5092-5097.

5. Allison, T. M., Hutton, R. D., Jiao, W., Gloyne, B. J., Nimmo, E. B., Jameson, G. B., and Parker, E. J. (2011) An extended β7α7 substrate-binding loop is essential for efficient catalysis by 3-deoxy-D-*manno*-octulosonate 8-phosphate synthase, *Biochemistry 50*, 9318-9327.

6. Jiao, W., Hutton, R. D., Cross, P. J., Jameson, G. B., and Parker, E. J. (2011) Dynamic cross-talk among remote binding sites: the molecular basis for unusual synergistic allostery, *J. Mol. Biol.* Doi:10.1016/j.jmb.2011.11.037.

Part A:

The calpain-calpastatin system

Chapter 1: Introduction to calpain and calpastatin

1.1 Calpain

Calpain is a cysteine protease that is responsible for catalysing the hydrolysis of the peptide bonds within a number of guest substrates. Substrates for calpain protease include calmodulin-binding proteins, the cytoskeletal proteins and the G-proteins (*1-6*). The active site of calpain contains a highly conserved catalytic triad, which is made up by Cys115, His272, and Asn296 (μ-calpain numbering). The sulphydryl group of Cys115 acts as the nucleophile facilitating the peptide bond cleavage.

Calpain requires the presence of calcium ions in order to be activated (*7*). The two major isoforms of calpain are the μ- and m-calpain (also named as calpain 1 and 2 respectively) and they mostly differ in calcium requirements for activation. μ-Calpain requires micro molar concentration of calcium ions to be activated, and m-calpain requires milli molar calcium concentration to be activated *in vitro*. Over-expression of calpain is associated with many diseases including brain trauma, spinal cord injury, Alzheimer's disease, Parkinson's disease, muscular dystrophy, arthritis, and cataract (*8, 9*). In recent years, considerable effort has gone into developing small molecule inhibitors that are potent and selective for calpain as the pharmaceutical target (*8*).

1.1.1 Structural features of human m-calpain

The crystal structure of the inactive calcium-free human m-calpain has been reported (*10*). The protein consists of six domains (Figure 1. 1), and these domains can be grouped into two subunits. The large subunit contains Domain I (DI) to Domain IV (DIV), and the small subunit contains Domain V (DV) and Domain VI (DVI). Domain II (DII) is the catalytic domain of calpain, and contains the active site residues. DII can be further subdivided

into two subdomains, Domain IIa (DIIa) and Domain IIb (DIIb) and both these subdomains contribute to the active site. DI is embedded in DVI of the small subunit, helping hold the two subunits of the calpain molecule together in the inactive form. DIV and DVI are the calcium binding domains. They each contain five EF-hands, which are typically helix-loop-helix sub-structures. Four of the five EF-hands can bind to calcium ions, and the fifth EF-hand in each of DIV and DVI interacts with each other to form the heterodimer interface between the large and small subunits of calpain.

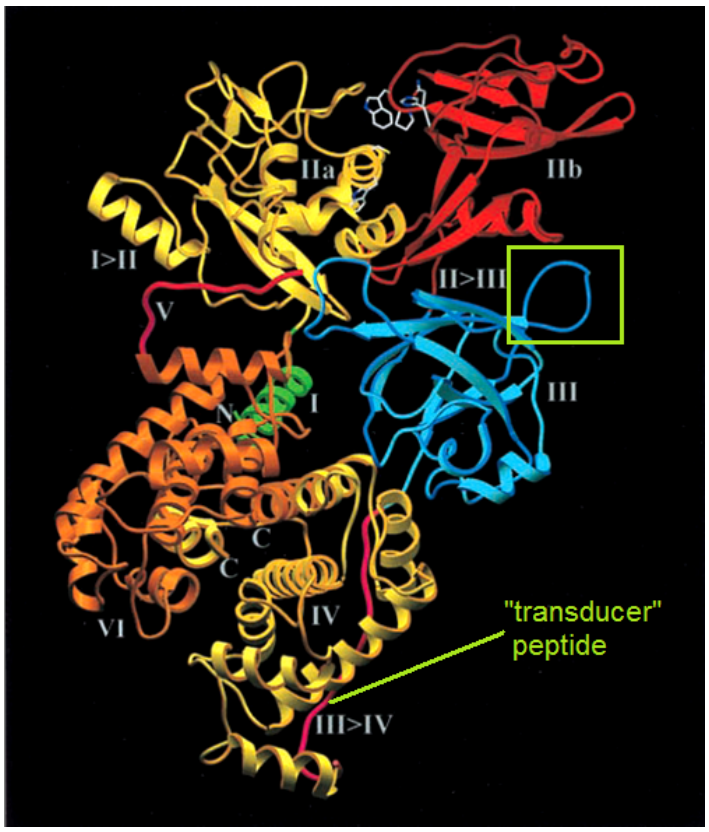


Figure 1. 1. The crystal structure of calcium-free human m-calpain (*10*).

DIIb is connected to DIII, and the short helix near the end of DIIb interacts with the acidic loop in DIII (highlighted in the green box in Figure 1. 1). The acidic loop contains many acidic residues, which normally interact with the basic residues in the nearby helix in DIIb (Figure 1. 2). These salt bridge interactions can help in holding DIIb away from DIIa, and keep calpain in its inactive form.

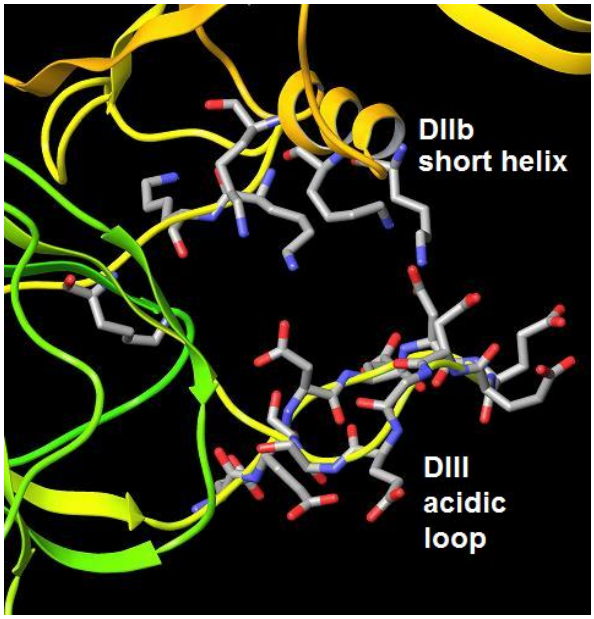


Figure 1. 2. Interactions between DIIb short helix (rich in basic Lys residues) and DIII acidic loop (rich in acidic Glu and Asp residues).

1.1.2 Activation of calpain in the presence of calcium ions

DII, DIV and DVI all can bind to calcium ions when they are present. The binding of calcium ions causes an overall conformational change in the molecule which leads to the activation of calpain. Firstly the binding of calcium ions to DVI causes conformational changes within this domain, and this leads to the release of DI helix and DIIa from DVI facilitated by autolysis[1] of DI. Secondly the associative interaction between the DIII acidic loop and DIIb short helix is broken due to binding of calcium ion in DIV, because the conformational changes in DIV leads to pulling of the "transducer" peptide (Figure 1. 1) that connects DIVand DIII, and hence helps the separation of DIII away from DIIb (*11, 12*).

After a series of conformational changes that occur within the calpain molecule upon binding of calcium ions, the two subdomains of DII, DIIa and DIIb, are released from the structural constraints that were holding them away from each other. The two subdomains then move closer to form the complete active site (Figure 1. 3).

[1] DI is cut off from the attached calpain molecule by another calpain molecule.
[2] The amino acid sequence for this 27-mer peptide is: DPMSSTYIEELGKREVTIPPKYRELLA

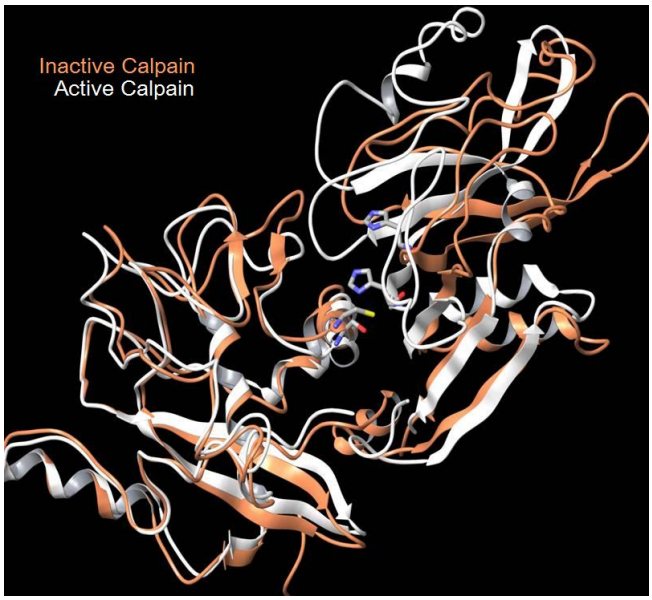


Figure 1. 3. Superposition of calpain DII in its active and inactive form. The active calpain catalytic domain is taken from the crystal structure 1KXR (rat μ-calpain) and the inactive calpain catalytic domain is taken from the crystal structure 1MDW (rat m-calpain). The two subdomains DIIa and DIIb are closer to each other in the active form than in the inactive form.

1.1.3 Active site of calpain

The active site of calpain consists of a highly conserved catalytic triad, which includes the active site cysteine, histidine and an asparagine residue (Figure 1. 4). When calpain is in the inactive form, the two catalytic subdomains are apart from each other, and the distance between the active site cysteine and histidine is about 8 Å (Figure 1. 4). After activation by calcium ions, the two subdomains close up and complete the active site, and the distance between the active site cysteine and histidine becomes much closer, about 3 Å ~ 4 Å (Figure 1. 4). The close distance allows for the histidine side chain to deprotonate the cysteine sulphydryl group, assisting its nucleophilic role in the hydrolysis of an encapsulated peptide bond of the substrate.

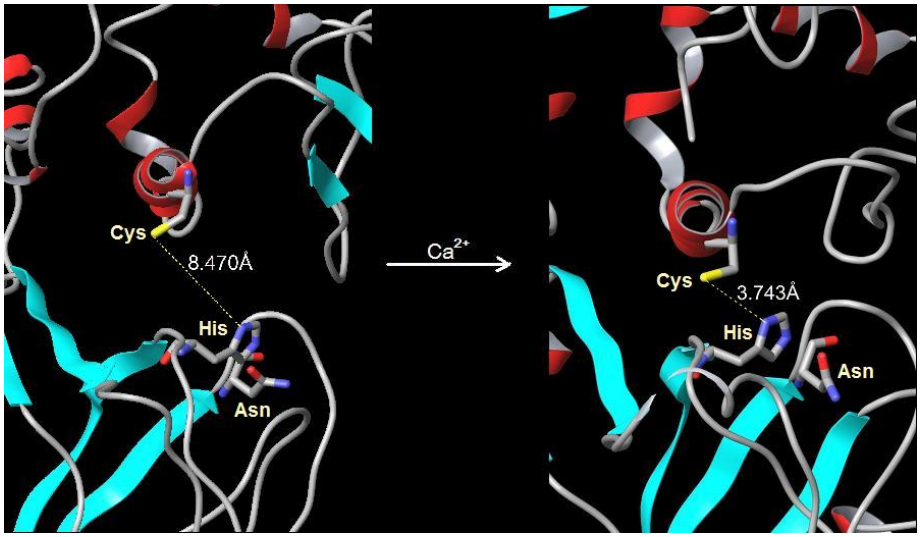


Figure 1. 4. The catalytic triad of calpain before and after activation by calcium.

1.2 Calpastatin

Calpastatin is the naturally occurring regulator of calpain. It co-exists with calpain in the cell and interacts specifically with calpain but no other cysteine proteases. The amino acid sequence is unique; there is no homology in amino acid sequence of calpastatin with any other proteins known to date (*13*). The calpastatin sequence is rich in polar and charged residues, and poor in hydrophobic residues.

1.2.1 Amino acid sequences of calpastatin

Amino acid sequences of mammalian calpastatins have been determined by cDNA cloning. The deduced amino acid sequences of human, pig, rabbit, cow, and sheep calpastatins consist of 708, 713, 718, 705, and 723 residues, respectively (*14-18*). Human calpastatin has five domains (shown schematically in Figure 1. 5). The L domain is at the N-terminus. Biological assays of the recombinant protein composed of the domain L show that this domain does not inhibit either μ- and m-calpain (*19-21*). Domains 1 to 4 are four repetitive inhibitory domains and each contains three highly conserved regions A, B and C. The calpastatin molecule contains little secondary

structure, with only regions A and C in each inhibitory domain having a tendency to form α-helical structures (*22*).

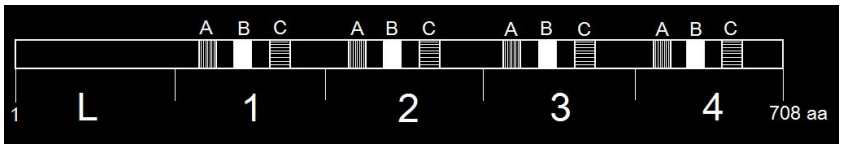


Figure 1. 5. Schematic structure of human calpastatin.

1.2.2 Regions A and C of the inhibitory domains of calpastatin

The four inhibitory domains of calpastatin are repetitive and mutually homologous, each containing about 140 amino acid residues. These four domains have been individually expressed and assayed against calpain and have been shown to inhibit calpain independently (*23-25*). Therefore, one calpastatin molecule can bind four calpain molecules corresponding to the number of inhibitory domains in calpastatin. There are three highly conserved regions in each inhibitory domain, called A, B, and C. Regions A and C are found to bind to calpain DIV and DVI respectively, and region B binds in the active site to inhibit calpain (*26, 27*).

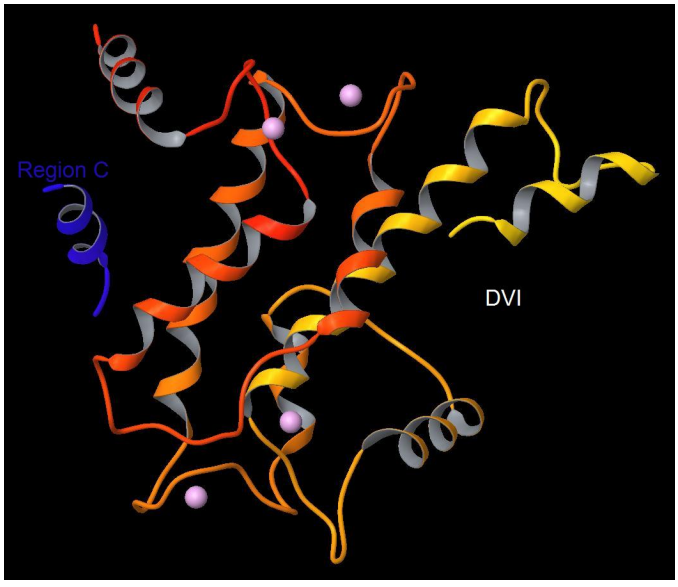


Figure 1. 6. Crystal structure of calpastatin inhibitory Domain 1 region C binding to calpain DVI (PDB 1NX1).

The crystal structure of calpain DVI interacting with a fragment of

calpastatin corresponding to region C of inhibitory Domain 1 has been reported (*26*). Calpain DVI binds four calcium ions with four of its EF hands. Region C adopts a helical structure upon binding to calpain (Figure 1. 6). The side chains of the hydrophobic residues Phe, Leu, Ile, and Ala in the calpastatin fragment interact with the corresponding binding pockets formed in DVI of calpain when calcium ions are bound (Figure 1. 7).

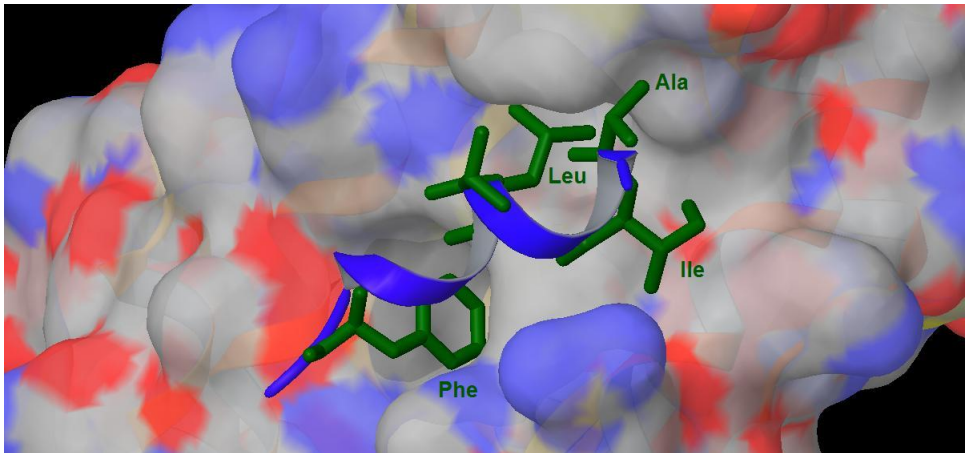


Figure 1. 7. Crystal structure of calpastatin inhibitory Domain 1 region C (shown in stick representation) binding to calpain DVI with surface of DVI displayed.

The interactions between region C of calpastatin and DVI of calpain have been studied with crystal structure data, from which these interactions can be examined at the molecular level. DIV, the calcium binding domain in calpain large subunit is very similar to DVI, and is therefore expected to show similar type of interactions with region A of calpastatin inhibitory domain (*26*).

1.2.3 Region B of calpastatin inhibitory domain

In 1999, Ma *et al.* studied the functions of the three conserved regions (A, B and C) in calpastatin inhibitory domain using site directed mutagenesis (*28*), and they demonstrated that when region B is completely removed from the calpastatin fragment, the inhibition activity of calpastatin is completely eliminated, while the binding affinity for calpain DIV is still retained. It is also shown that region B on its own still retains half of the calpain inhibition activity although it does not bind to DIV of calpain. This result clearly shows

that region B of the calpastatin inhibitory domain is essential for inhibition and regions A and C can help to increase inhibition efficiency by binding to calpain regulatory domains and bring region B into close proximity to calpain active site.

1.3 Overall interactions between calpastatin and calpain

In 2008, Moldoveanu *et al*. and Hanna *et al*. reported two crystal structures of the calpain-calpastatin complex (*29, 30*). For the first time, the overall binding mode of calpastatin on calpain is revealed at atomic level detail. These two structures also revealed the binding site and binding mode of the most critical region in calpastatin, region B, which could only be speculated before (Figure 1. 8).

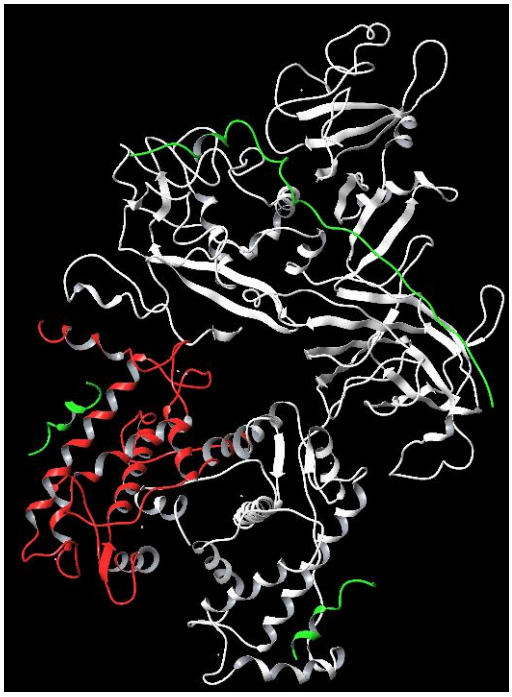


Figure 1. 8. Calcium-dependent complex between m-calpain and calpastatin (PDB 3DF0). DII, DIII and DIV of the calpain large subunit are displayed in white; the calpain small subunit is displayed in red and the calpastatin molecule (regions A, B, and C) is displayed in green.

The mode of calpastatin action found in the crystal structure is consistent with what was proposed from previous experimental studies (Figure 1. 8).

The regions A and C of calpastatin inhibitory domain adopt α-helical structure and bind to DIV and DVI of calpain respectively. Region B which is critical for inhibition binds to calpain to block access into the active site. Half of the inhibitory region B binds inside the calpain active site cleft, while the other half interacts with calpain DIII.

1.4 Part A of this thesis

Part A of this thesis describes computational studies conducted for the calpain-calpastatin system. In particular, the dynamic conformational preferences of region B of inhibitory domain in calpastatin were examined in detail in Chapter 2. The knowledge gained in these studies can be used to help in the design of calpain inhibitors constrained to adopt the preferred inhibitory conformation and thereby achieve compounds with heightened potency and selectivity.

Chapter 2: Calpastatin inhibitory domain region B: conformational features of the CP1B peptide

2.1 Introduction

In 1989, Maki *et al.* found that a synthesised 27-mer peptide[2, 3] containing the central conserved sequence of region B of Domain 1 of human calpastatin, possesses potent inhibitory activity against calpain (*16*). This peptide is specific for calpain and shows no inhibition activity against papain or trypsin. In 1991, Ishima *et al.* examined the solution structure of this 27-mer peptide in DMSO-d_6 by two-dimensional ^{1}H NMR spectroscopy and found no evidence for any regular α-helix or β-sheet (*31*). However, they reported the possibility for a turn in the region from Glu10 to Lys13, which contains a highly conserved residue Gly12, in all calpastatins from different species and in the four repetitive inhibitory domains of each calpastatin. These findings suggest that although the repetitive domain of calpastatin does not have an extensive tertiary structure, it has a well-defined local turn structure in the well conserved region which is considered essential for the specific interaction with calpain.

In 2003, Anagli *et al.* used a combination of β-alanine scanning mutagenesis and kinetic measurements of the binding of this 27-mer peptide with calpain to probe the relative contributions of each of the amino acid side chains and the backbones to the overall calpain-inhibitory activity (*32*). The study identified two "hot spots", Leu11-Gly12 and Thr17-Ile18-Pro19, in the peptide which contain residues critical for inhibitory function. Mutation of any one of the key residues in either of the two hot spots resulted in a dramatic loss of inhibitory activity. The individual contributions of the amino acid residues to the overall inhibitory activity were ranked from this study:

[2] The amino acid sequence for this 27-mer peptide is: DPMSSTYIEELGKREVTIPPKYRELLA
[3] Amino acid one-letter codes: Gly(G), Ala(A), Ser(S), Thr(T), Cys(C), Val(V), Leu(L), Ile(I), Met(M), Pro(P), Phe(F), Tyr(Y), Trp(W), Asp(D), Glu(E), Asn(N), Gln(Q), His(H), Lys(K), and Arg(R).

Ile18 > Leu11 > Gly12 > Pro19 > Thr17 > Arg14 > Glu10 > Glu9 >= Pro20 > Tyr7 > Tyr22 > Asp1 >= Lys13 > Val16 > Leu26 > Leu25 > Ile8 > Ala27 > Thr6 > Lys21 >= Glu15 > Arg23 > Met3 > Pro2 >= Ser5 > Glu24 > Ser4.

In 2004, Anagli *et al*. analysed the contributions of the side chains of Leu11 and Ile18 in even more detail (*33*), by using mutagenesis techniques, and found that β- and γ-CH_2 of the Leu11 and Ile18 side chains respectively play a direct role in calpain inhibition.

A 20-mer peptide with amino acid sequence SSTYIEELGKREVTIPPKYR (sequence numbering starts at 4 and finishes at 23, to allow consistency with sequence numbering in the 27-mer peptide), identified by systematic truncation of the 27-mer peptide of calpastatin region B, has been reported to be the core sequence required to maintain affinity and selectivity towards calpain (*34*). This 20-mer peptide was named CP1B to depict the region of calpastatin it was derived from.

The importance of the turn-like region between Glu10 and Lys13, that was identified earlier from the NMR studies of the 27-mer, has been investigated by a study of a series of conformationally constrained macrocyclic variants of CP1B (*34*). These macrocyclic compounds were assayed against calpain, and the results showed one of these (Figure 2. 1) was as potent as the 'linear' CP1B peptide. Modelling showed that this macrocycle contains a loop-like conformation in the region from amino acid 10 to 13. The high inhibitory activity of this macrocycle suggests that the activity of the linear peptide CP1B is a consequence of it adopting a similar loop-like conformation in the region between Glu10 and Lys13.

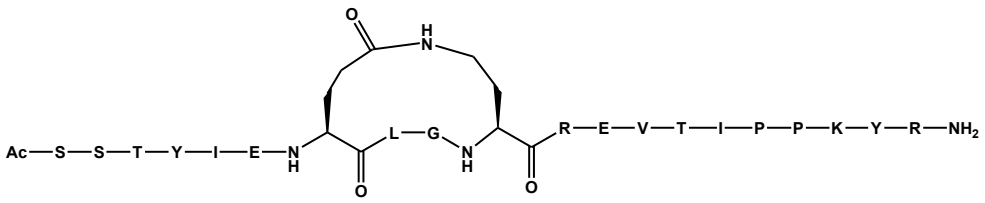


Figure 2. 1. Structure of the macrocyclic compound that is as potent as linear CP1B peptide.

The Pfizer group postulated that with the formation of this loop, the two hydrophobic residues Leu11 and Ile18 of the CP1B peptide are positioned to interact with their corresponding pockets on calpain (*34*). The group also proposed that it is the presence of oppositely charged residues at positions 10 and 13 on either side of the loop region that favours the formation of the loop conformation, because they can form salt bridges which stabilise the loop conformation of the peptide.

Two crystal structures of the calpain-calpastatin complex were released in 2008 by Moldoveanu *et al*. and Hanna *et al*. and showed in detail the interactions between region B of calpastatin and calpain (*29*), (*30*). The section of region B of calpastatin, which corresponds to the 20-mer peptide CP1B, is found to bind within the active site cleft of calpain. One of the two crystal structures (3DF0) (Figure 2. 2) shows the active site of calpain in complex with calpastatin region B (*30*).

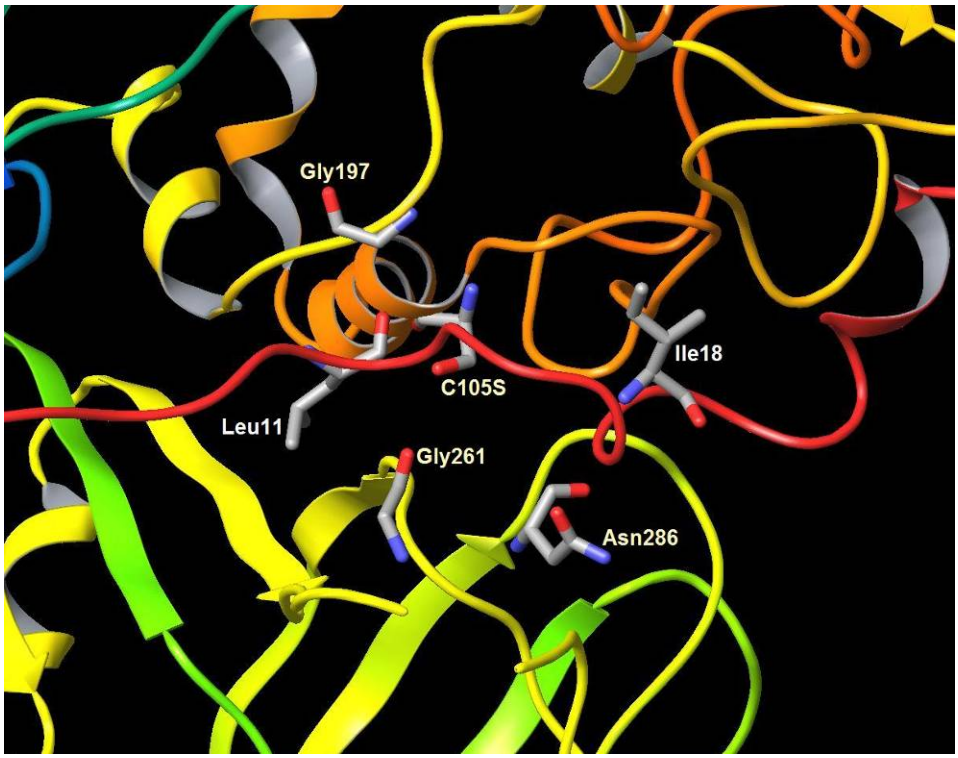


Figure 2. 2. Active site of calpain in complex with calpastatin (shown in red) (PDB 3DF0). The key residues in both calpain active site and region B of calpastatin are displayed with tube representations.

The two critical hydrophobic residues, Leu11 and Ile18, on calpastatin interact with two corresponding hydrophobic pockets in the calpain active site. Ile18 interacts in the prime site of calpain active site (right hand side of the active site Cys105S in Figure 2. 2) and Leu11 in the non-prime site (left hand side of Cys105S in Figure 2. 2). The binding pockets in calpain are more clearly shown in Figure 2. 3, where Leu11 is shown interacting with the S2 binding pocket of calpain and Ile18 is interacting with the S2' pocket. These two hydrophobic interactions together with other non-covalent interactions help to keep the calpastatin molecule in the active site and exclude other molecules accessing the calpain active site thereby exerting an inhibitory action.

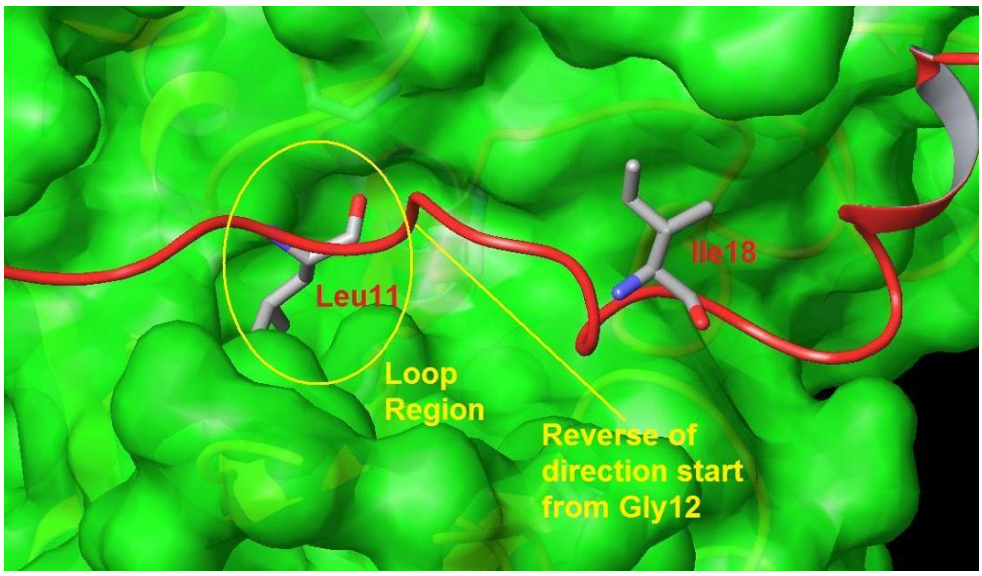


Figure 2. 3. Interaction between calpastatin inhibitory region residues Leu11 and Il18 and calpain active site.

The loop region first proposed by Pfizer *et al.* is highlighted in yellow circle in Figure 2. 3. To the right of this loop is a second loop adjacent to Leu11 that had not been predicted. In the crystal structure the peptide adopts an overall S-shape in the active site, where the second reverse in direction starts with the amino acid Gly12 immediately adjacent to Leu11. The "kink" ensures that the peptide bond near the active site cysteine is moved away from this thiol nucleophile (Figure 2. 4) thereby preventing proteolysis[4] of

[4] Proteolysis is the directed degradation (*digestion*) of proteins by cellular enzymes called proteases or by intra-molecular digestion.

calpastatin region B by calpain (*30*). In this way, calpastatin acts to bind in the active site of calpain and prevents other molecules from accessing the active site, but for the conformational reasons referred to above is not digested by calpain. Moldoveanu *et al*. suggested that calpastatin is not digested and can therefore be recycled after being released from calpain.

The second loop region was not detected in the NMR studies (*31*) of calpastatin region B, or in the macrocyclic analogue study (*34*). The failure to observe a second loop in calpastatin region B in the enzyme-free state suggests that CP1B is prearranged to form only one loop at Leu11 of region B, and we propose that it could be upon binding to calpain that Ile18 residue moves to bind in S2' pocket of the active site and that this forces the second "kink" to be formed.

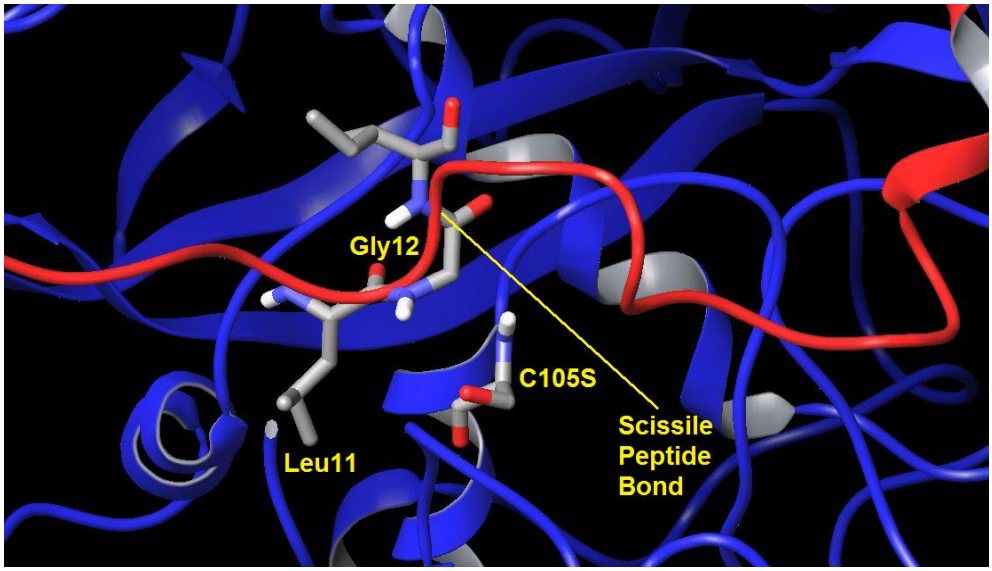


Figure 2. 4. The "kink" at Gly12 moves the scissile peptide bond in calpastatin away from the active site cysteine. Calpastatin is shown in red and calpain is shown in blue.

The aim of the work in this chapter is to identify the essential key structural features within CP1B that have a significant influence on the peptide's inhibition activity. This chapter describes a computational study of the 20-mer peptide CP1B and its various mutants in the absence of calpain, in order to establish the dynamic conformational preferences of this peptide and the specific contribution of each residue in dictating the pre-formed loop region around Leu11 in CP1B. The modelling techniques used for this study

include molecular dynamics (MD) simulations conducted with the software NAMD and stochastic dynamics simulations combined with Monte Carlo sampling method (MCSD). The modelling results were examined and compared with measured IC_{50} data. The results illuminate the importance of the prearrangement of this loop region in determining the inhibitory activity of CP1B.

2.2 Modelling methods

2.2.1 Molecular dynamics simulations

In a molecular dynamics (MD) simulation, the motion of the system in a certain time period is simulated according to the Newtonian equation of motion:

$$m_\alpha a_\alpha = F_\alpha \qquad (1)$$

where m_α is the mass of atom α, a_α is the acceleration of the atom α, and F_α is the force in the system that is acting on the atom α. The acceleration term a_α is calculated by second order differentiation of the positional vector of atom α, i.e.

$$a_\alpha = \partial^2(r_\alpha)/\partial t^2 \qquad (2)$$

where r_α is the positional vector of atom α, and t is the time from the start of the simulation. The force term F_α is calculated as the negative gradient of the total potential energy which is dependent on all atomic positions, i.e.

$$F_\alpha = -\partial U_{total}/\partial r_\alpha \qquad (3)$$

where U_{total} is the total potential energy term, and by substituting equations (2) and (3) into equation (1), Equation (4) can be obtained:

$$m_\alpha \partial^2(r_\alpha)/\partial t^2 = -\partial U_{total}/\partial r_\alpha \qquad (4)$$

Equation (4) is the Newtonian equation in terms of time and atomic positions and this is the form of the equation that is used in a MD simulation.

The most important term in Equation (4) is the total potential energy term, U_{total}. This potential energy is represented through the MD *force field*. A

force field includes all types of atomic interactions in a function that is in simple mathematical form and therefore can be calculated quickly. The force field function used to represent the total potential energy includes the following contributions:

$$U_{total} = U_{bond} + U_{angle} + U_{dihedral} + U_{vdw} + U_{Coulomb} \qquad (5)$$

The first three terms in Equation (5) describe the bonded interactions such as stretching, bending and torsional interactions. The final two terms in Equation (5) describe interactions between non-bonded atoms, where U_{vdw} corresponds to van der Waal's interactions approximated by a Lennard-Jones 6-12 potential and $U_{Coulomb}$ corresponds to electrostatic interactions. Each of the five potential energy terms is written as the function of the positional vectors r_i, and they are listed below:

$$U_{bond} = \sum_{bonds\ i} k_i^{bond}(r_i - r_{0i})^2$$

$$U_{angle} = \sum_{angles\ i} k_i^{angle}(\theta_i - \theta_{0i})^2$$

$$U_{dihedral} = \sum_{dihedral\ i} k_i^{dihe}[1 + \cos(n_i\emptyset_i - \gamma_i)]$$

$$U_{vdw} = \sum_i \sum_{j>i} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$

$$U_{Coulomb} = \sum_i \sum_{j>i} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$

In the above equations, r in U_{bond} represents bond length; θrepresents bond angle term; $\emptyset$ represents the dihedral angle term; r_{ij} terms in U_{vdw} and $U_{Coulomb}$ represent distance between pairs of non-bonded atoms. The parameters such as force constants k_i and bond length r_{0i} etc used in the above 5 equations for calculating the five different energy terms are listed in the force field *parameter* files. These parameters are determined through a combination of empirical techniques and quantum mechanical calculations

and then tested for reliability using small systems that are well characterised.

Periodic boundary conditions (PBC) are used in MD simulations to make sure the system does not have an abrupt border with the vacuum. In PBC, the modelled system (including solvent surround the biological system) is placed in a *unit cell*. This unit cell is considered to have infinite number of images in space. Any atom that leaves the unit cell on one side is replaced by its image entering on the other side of the cell.

The van der Waal's interactions and the Coulomb electrostatic interactions exist between all non-bonded atom pairs, even when the atoms are far away from each other. Therefore calculating every long range interactions exactly is not practical. In MD simulations conducted with the software NAMD, the long range van der Waal's interactions are spatially truncated at a cut-off distance of 25Å. The full long range electrostatic interactions are computed by making use of the periodicity of the periodic boundary condition and combined with the particle-mesh Ewald (PME) method (*35*).

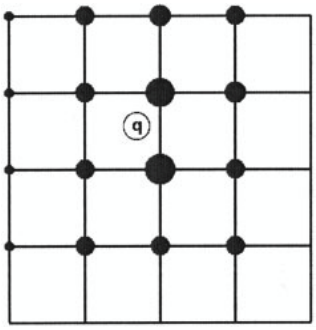


Figure 2. 5. Distribution of charge q over the grid (shown in 2D mesh here) (*36*).

In PME, the total electrostatic interaction energy is split into two summations: the direct sum of short range electrostatic interactions which converges quickly in real space; and the Ewald summation (*37*) of the long range interactions which converges quickly in Fourier space. The Ewald sum of long range interactions is basically the sum of the Fourier transforms of charge density and potential. In order to evaluate the Fourier transform of the charge density, the Fast Fourier Transform (FFT) is used. Since FFT requires the charge density to be evaluated on a discrete lattice, the charge is

distributed over three dimensional grids (Figure 2. 5), hence the name "mesh" in the PME method.

2.2.2 Stochastic dynamics combined with Monte Carlo sampling

In stochastic dynamics simulations the solvent molecules are not modelled explicitly as in MD. Instead the effect of the surrounding solvent is modelled by introducing a random force combined with a damping force.

The use of Monte Carlo sampling can introduce large changes into a system during a simulation, including changes in peptide backbone torsion angle. Therefore it is more efficient in sampling conformational space of a molecule than MD method, where the molecule can be trapped in a local minimum for a long time.

In this chapter, a series of simulations for the overall conformations of both CP1B wild type (WT) and its various mutants was carried out using the method MCSD, which was the combined method of stochastic dynamics and Monte Carlo sampling. The combined method can sample the local minimum quite efficiently and also can explore the total conformational space of the peptide.

2.2.3 Simulations set up in this chapter

2.2.3.1 Set up of MCSD simulations in this chapter

MCSD simulations for CP1B WT and selected β-Ala mutants were carried out with the modelling software package MacroModel using the OPLS2005 force field with the GB/SA water model to mimic solvent effect (*38*).

For each mutant and the WT system of CP1B, three repetitive (50 ns each) simulations were carried out starting with similar linear conformations. The starting linear conformation of the WT peptide CP1B was established using the protein builder in the Schrödinger GUI Maestro (*39*). The analogous

starting conformations for the various mutants of CP1B were generated by mutating corresponding residues on CP1B WT into β-Ala *in silico* in line with the experimental changes made to the peptide by Betts *et al* (*32*).

For CP1B WT, two other MCSD simulations (50 ns each) were carried out starting with half-folded and fully-folded conformations respectively. The half-folded starting conformation of CP1B WT was the conformation in the last frame of a 1 ns MD simulation for CP1B. The starting fully-folded conformation used in these calculations was the lowest energy conformer obtained from a conformational search for the WT peptide. This conformational search was conducted with the MCMM Serial Torsional Sampling method and was run with a GB/SA water model, using the OPLS2005 force field, with 3000 steps for the conformational search and an energy window of 12.0 kJ/mol for collecting appropriate conformers.

For all MCSD simulations conducted in this study, there was no cut-off distance applied for electronic and van der Waals interactions. The starting structure in each case was first minimised for up to 5000 steps with a gradient convergence threshold of 0.05. Stochastic dynamics was then carried out for 50 ns with the SHAKE procedure applied to all bonds to hydrogen atoms. The simulations were run at a constant temperature of 300K and with a time step of 1.5 fs (*40*). During the dynamics simulation the torsions within the peptide were subject to changes by the application of the Monte Carlo method. Five hundred structures were sampled in equal time intervals for later examination and analysis.

2.2.3.2 Set up of MD simulations in this chapter

To confirm that the conformations obtained from MCSD simulations were reasonable, three 100 ns long MD simulations for CP1B WT were carried out with the software NAMD running on the BlueFern™ supercomputer at the University of Canterbury (*36*). The starting conformations of these three MD simulations were obtained from the final frames in the three MCSD simulations for CP1B WT starting from linear, half-folded and folded

conformations as described above. The peptides were solvated with explicit water molecules in a water box. The water box in each system contains approximately 1500~1600 water molecules. The simulations were carried out with CHARMM force field parameter specifications at constant temperature and pressure at 310 K and 1 atmosphere (*41*). The cut-off distance for van der Waals interactions was set to be 12 Å. In each simulation, the system was first minimised for 5000 steps followed by 100 ns of dynamics simulation carried out at 2 fs time steps. The trajectory was written out at 100 ps time interval, and a total of 1000 frames were obtained from each simulation (100 ns). These full-atom MD simulations are significantly more computationally expensive than the MCSD method. With NAMD using 32 processors running on the BlueFern^{TM} supercomputer, each of the three 100 ns MD simulations takes 3~5 weeks to complete. A sample of the job setup script for MD simulations with NAMD is in appendix I.

2.2.3.3 Cluster analysis of sampled conformations

Cluster analyses were conducted on the sampled structures using XCluster (*42*) in order to examine the conformations explored during the MCSD and MD simulations. The clustering was performed in a hierarchical fashion, based on the pairwise RMS displacements of selected comparison atoms. In this study, the comparison atoms were set to be the backbone heavy atom of the loop region Glu10-Lys13 for CP1B WT system, and the region Glu9-Glu15 for the β-Ala mutant systems. The clusters with the most members were considered to comprise the most stable conformations explored during the dynamics simulations, and the representative structures of the largest clusters found in these simulations were selected as follows. For each cluster, the centroid of each of the comparison atoms was determined. Then the structure within the cluster which had the minimum RMS inter-atomic distance from these centroids was selected as the most representative structure of this cluster. These representative structures were then used in further structural comparisons.

2.3 Results and Discussions

2.3.1 Dynamics of CP1B WT

To examine the conformational preference of the WT CP1B peptide, particularly the proposed loop region (from Glu10 to Lys13), in the absence of the calpain enzyme, MCSD simulations were conducted for the WT system.

Table 2. 1. Cluster analyses of sampled structures from the MCSD and MD simulations for CP1B WT.

Simulation ID	Simulation length	Simulation type	Number of clusters	Cluster sizes
1	50 ns	MCSD	2	270, 230
2	50 ns	MCSD	3	152, 1, 347
3	50 ns	MCSD	2	499, 1
4	100 ns	MD	3	35, 64, 1
5	100 ns	MD	4	9, 8, 8, 75
6	100 ns	MD	4	90, 1, 8, 1

Table 2. 2. Overall and loop region backbone RMSD values in the three 50 ns MCSD simulations for CP1B WT peptide. The RMSD values are measured between the starting conformation and the average structure of the largest cluster in each simulation.

Simulation ID	Simulation length	Starting conformation	Backbone RMSD (Å)	Loop region (Glu10-Lys13) backbone RMSD (Å)
1	50 ns	Linear	16.8	2.22
2	50 ns	Half-folded	10.7	1.97
3	50 ns	Folded	2.23	0.30

Three MCSD (50 ns) simulations of the CP1B WT peptide were initiated starting from linear, half-folded and fully-folded conformations respectively. Cluster analyses were carried out to examine the conformations explored during these simulations (Table 2. 1). In the cluster analyses, pairwise RMS displacements in backbone atoms of the loop region (Glu10-Lys13) were calculated, and the sampled structures were clustered based on these displacements. The largest cluster in each simulation was considered to

comprise the most stable conformations explored, and the representative structure of all members in the largest cluster was obtained to represent the cluster and to be used in further structural comparisons. The backbone RMSD values for the entire peptide and the loop region (Glu10-Lys13) between the representative structures and the starting conformations in these simulations are listed in

Table 2. **2**. RMSD measurements for two of the MCSD simulations (Simulations 1 and 2,

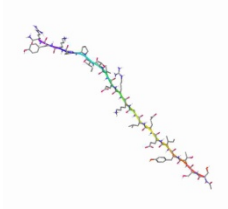
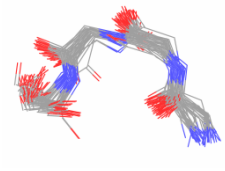
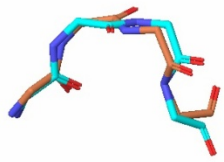
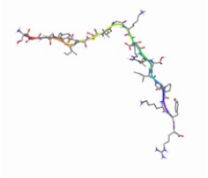
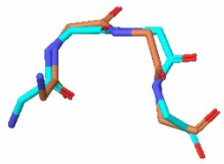
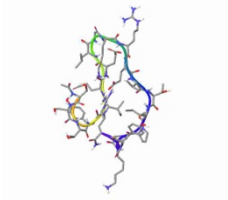
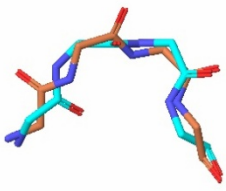
Table 2. **2**), starting from the linear and half-folded conformations, were large, indicating the MCSD method is sampling significant conformational space. Simulation 3 which started with folded conformation shows much smaller backbone changes. This observation is consistent with the folded starting conformation for this calculation representing a low energy conformer of the CP1B WT peptide. Therefore, this simulation started from a favourable low energy conformation and exhibited limited sampling outside this local minimum during the entire simulation.

To ensure efficient sampling over large conformational space, all further MCSD simulations for both CP1B WT and its mutants were set to begin with the extended linear conformations. The backbone RMSD values for the loop region (

Table 2. **2**) were generally found to be smaller than the overall backbone RMSD variation, but still gave the same trend, with Simulations 1 and 2 producing larger RMSD values than Simulation 3. Importantly all three simulations resulted in folded conformations (

Table 2. 3), demonstrating that the peptide has a strong preference for an overall folded loop-like conformation.

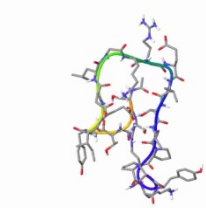
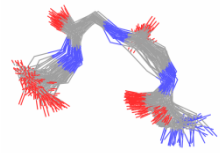
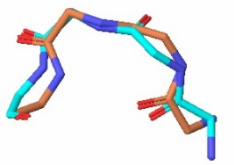
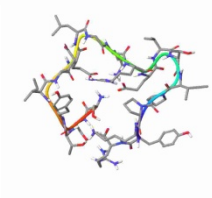
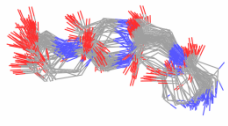
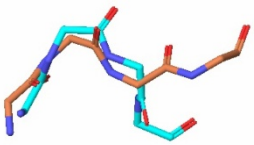
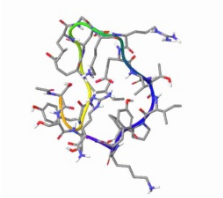
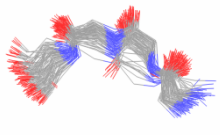
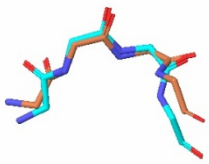
Table 2. 3. Table of starting conformations, simulation ensembles of the loop region backbone for CP1B WT peptide in the three 50 ns MCSD simulations. The superposition of the representative structure from each simulation (brown carbon) and the binding conformation in the crystal structure (cyan carbon) is also displayed.

Simulation ID	Starting conformation	Simulation Ensemble of the loop region backbone (collected every 1 ns)	Superposition of the loop region between the representative structures and the binding conformation
1. (Fully extended linear conformation)			
2 (Half folded conformation)			
3 (Fully folded conformation)			

The MCSD method treats solvent implicitly and combines stochastic dynamics with Monte Carlo sampling. To exclude the possibility that folding observed in simulations 1 and 2 were artefacts of this particular simulation

method, the conformations obtained from the last frames in these three MCSD simulations were used as the starting geometries for three more rigorous 100 ns MD simulations, in which the solvent was treated explicitly. This was done to establish if the loop region between Glu10 and Lys13 would start to unfold.

Table 2. 4. Table of starting conformations, simulation ensembles of the loop region backbone in the three 100 ns MD simulations for CP1B WT started from last frame conformations obtained from corresponding MCSD simulations. The superposition of the representative structures from each simulation (brown carbon) with the starting conformations (cyan carbon) at the loop region backbone is also displayed.

Simulation ID	Starting conformation	Simulation Ensemble of the loop region backbone (collected every 1 ns)	Superposition of the loop region between the representative structures and the starting conformations
4			
5			
6			

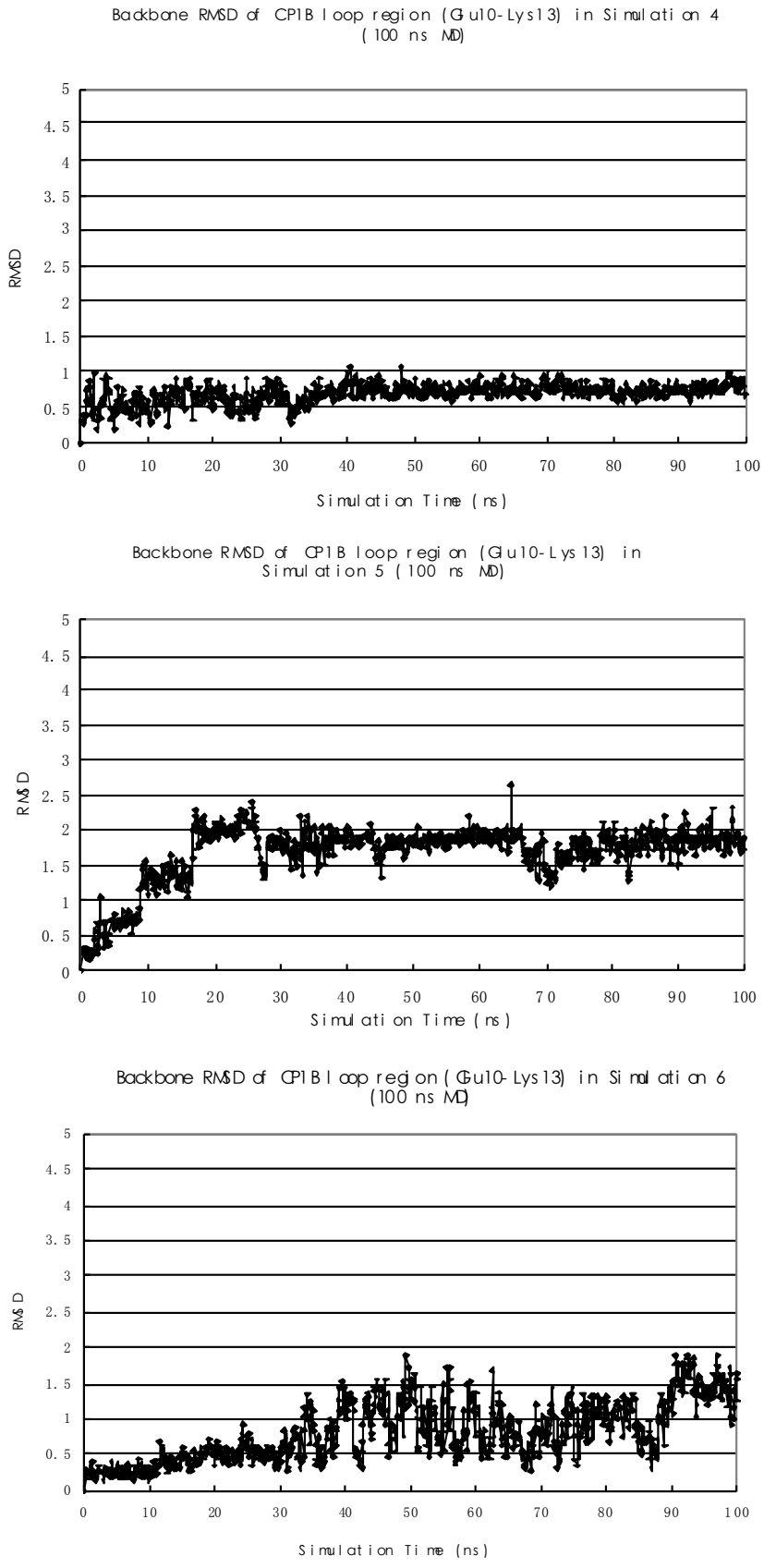


Figure 2. 6. Plots of backbone RMSD of the loop region (Glu10-Lys13) in the three 100 ns MD simulations (Simulations 4-6).

Examination of the simulation ensembles showed that the loop region (Glu10-Lys13) of WT CP1B remained folded during these three 100 ns MD simulations (

Table 2. **4**). The conformation of the loop region was generally observed to become stable after 35 ns into the MD simulation (Figure 2. 6). The same cluster analyses as described above were carried out to examine the sampled structures from these three MD simulations (Table 2. 1), and the representative structures of the largest clusters were obtained for structural comparison (

Table 2. **4**). The RMSD values of the loop region backbone were reasonably small (within 0.8 Å for Simulations 4 and 6, and slightly larger in Simulation 5, with a RMSD value of 1.84 Å) considering the system under study was a small flexible peptide (Table 2. 5). The RMSD values suggest that both the full-atom MD and the MCSD simulations are sampling in the same family of related loop conformations for the Glu10-Lys13 region. This observation validates the use of the less computationally expensive MCSD method for the following study, and furthermore confirms that the peptide indeed has an inherent propensity to fold.

Table 2. 5. Backbone RMSD values of the loop region in the three 100 ns MD simulations. The RMSD values are measured between the starting conformation and the average structure of the largest cluster in each simulation.

Simulation ID	Simulation length	Starting conformation	Loop region (Glu10-Lys13) backbone RMSD
4	100 ns	Last frame structure of Simulation 1	0.73
5	100 ns	Last frame structure of Simulation 2	1.84
6	100 ns	Last frame structure of Simulation 3	0.74

The modelling results obtained for the WT CP1B can be compared with two crystal structures of the calpain-calpastatin complex recently reported by Moldoveanu *et al*. and Hanna *et al*., which illustrate the interactions between region B of calpastatin and the active site of calpain at atomic level (*29, 30*).

In these two crystal structures, the section of region B of calpastatin that corresponds to the 20-mer peptide CP1B was found to bind within the active site cleft of calpain (Figure 2. 3).

The binding conformation of the loop region of CP1B in the crystal structure (3DF0) was used as the reference structure, and the backbone RMSD values for the loop region (Glu10-Lys13) were measured for all representative structures in the three MCSD simulations for CP1B WT (

Table 2. **6**,

Table 2. 3). All three simulations showed reasonably small backbone RMSD values for the loop region, with Simulation 1 displaying the smallest RMSD value (0.75 Å). The comparison between the crystal structure conformation and our modelling results suggests that even in the absence of the calpain enzyme, the region Glu10-Lys13 of WT CP1B has a propensity to fold into loop-like conformations which is very similar to the conformation known to bind.

Table 2. 6. Backbone RMSD of the loop region (Glu10-Lys13) between the binding conformation of CP1B in crystal structure (3DF0) and the average structures in the largest clusters of each MCSD simulation for CP1B WT.

Simulation ID	Starting conformation	Loop region (Glu10-Lys13) backbone RMSD compared to binding conformation (Å)
1	Linear	0.75
2	Half-folded	1.17
3	Folded	0.79

2.3.2 Dynamics of selected β-Ala mutants of CP1B and analysis

Following on from the studies of WT CP1B, studies were undertaken to establish the effect of mutations along the peptide on the conformational preference of this loop region in the absence of calpain. A prior study reported experimental IC_{50} values of various β-Ala mutants of the CP1B peptide (*32*). Therefore we elected to carry out β-Ala mutation *in silico* along the CP1B peptide in order to provide direct comparison with the reported experimental IC_{50} values.

A series of β-Ala mutants of CP1B was obtained by selective mutation of residues Glu9, Glu10, Gly12, Lys13, Arg14 and Glu15, near the proposed loop region *in silico*. The residue Leu11 was omitted from this study because it has been shown that the importance of this residue in calpain inhibition lies in its side chain functionality (*33*). The side chain of Leu11 acts to fit in one hydrophobic pocket in the calpain active site, and losing this interaction (by β-Ala mutation) dramatically decreases the inhibition of calpain by the CP1B peptide. Therefore, the effect of Leu11βAla mutation on the measured IC_{50} value is mainly contributed by the loss of its side chain functionality, rather than its possible effect on the conformation of this loop region. Unlike Leu11, the other residues (Glu9-Glu10, Gly12-Glu15) are likely to play a role in facilitating the formation of the loop-like conformation (*34*). The oppositely charged residues on either side of the loop region (Glu9, Glu10, Lys13 and Arg14) can form salt-bridges which would bring the two ends of the loop region into close proximity, and the highly flexible Gly12 can act to promote the formation of the loop.

Table 2. 7. Cluster analyses of sampled structures from MCSD simulations for various β-Ala mutants of CP1B.

Species	Simulations	Number of Clusters	Cluster sizes
Glu9BAla	Glu9BAla_1	2	1, 499
	Glu9BAla_2	2	499, 1
	Glu9BAla_3	3	1, 3, 496
Glu10BAla	Glu10BAla_1	5	2, 1, 72, 1, 424
	Glu10BAla_2	2	499, 1
	Glu10BAla_3	2	8, 492
Gly12BAla	Gly12BAla_1	2	1,499

	Gly12BAla_2	2	498, 2
	Gly12BAla_3	4	496, 2, 1, 1
Lys13BAla	Lys13BAla_1	3	14, 485, 1
	Lys13BAla_2	4	495, 2, 2, 1
	Lys13BAla_3	3	496, 2, 2
Arg14BAla	Arg14BAla_1	2	499, 1
	Arg14BAla_2	4	1, 1, 9, 489
	Arg14BAla_3	2	1, 499
Glu15BAla	Glu15BAla_1	3	2, 22, 476
	Glu15BAla_2	4	34, 464, 1, 1
	Glu15BAla_3	2	499, 1

For each β-Ala mutant, three repetitive MCSD simulations (50 ns) were initiated from similar linear conformations in order to enhance the likelihood that the conformational space is appropriately traversed. All MCSD simulations for all six β-Ala mutants produced folded loop-like conformations. A cluster analysis was carried out for each simulation to examine the conformations explored (Table 2. 7). In these cluster analyses, the pairwise backbone displacements in the region from Glu9 to Glu15 were calculated between all sampled structures in each simulation. The largest cluster was identified for each simulation and the representative structure was obtained to be used in the following structural comparisons. RMSD analysis was then used to examine the effect of these β-Ala mutations on the conformational preference of the loop region of CP1B. To ensure that the effect of any particular mutation was not over-represented, RMSD values are measured for the region from Glu9 to Glu15, i.e. all the mutated residues are included in RMSD calculations. Backbone heavy atom RMSD values for the region Glu9-Glu15 were measured between the binding conformation of CP1B in the crystal structure (3DF0) and the representative structures obtained from cluster analysis for the simulation ensembles of β-Ala mutants of CP1B (Table 2. 8,

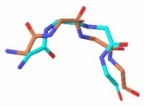
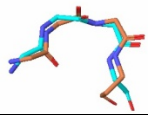
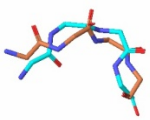
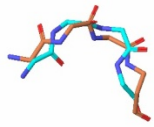
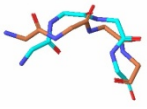
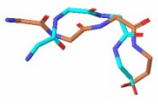
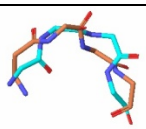
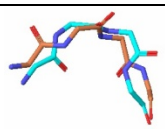
Table **2. 9**).

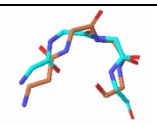
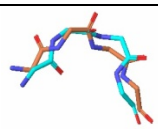
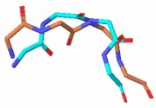
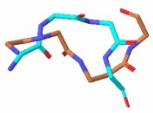
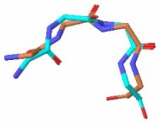
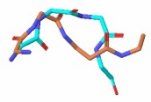
Table 2. 8. Backbone RMSD values for the region Glu9-Glu15 between the binding conformation of CP1B in the crystal structure (3DF0) and the average structure obtained from each MCSD simulations for the β-Ala mutants of CP1B. "Average RMSD" is the average RMSD values in the three repetitive MCSD simulations for each mutant

system.

Species	Simulations	Backbone RMSD values between average structure and the binding conformation	Average RMSD
Glu9BAla	Glu9BAla_1	2.80	3.58
	Glu9BAla_2	3.82	
	Glu9BAla_3	4.12	
Glu10BAla	Glu10BAla_1	4.76	4.88
	Glu10BAla_2	5.26	
	Glu10BAla_3	4.63	
Gly12BAla	Gly12BAla_1	5.68	5.43
	Gly12BAla_2	5.83	
	Gly12BAla_3	4.78	
Lys13BAla	Lys13BAla_1	3.09	3.36
	Lys13BAla_2	3.43	
	Lys13BAla_3	3.57	
Arg14BAla	Arg14BAla_1	4.08	3.13
	Arg14BAla_2	3.42	
	Arg14BAla_3	1.89	
Glu15BAla	Glu15BAla_1	3.29	3.10
	Glu15BAla_2	2.17	
	Glu15BAla_3	3.85	

Table 2. 9. Superposition of the loop region backbone between the representative conformation (brown carbon), obtained from each of the 50 ns MCSD simulations for the six β-Ala mutants of CP1B, and the binding conformation (cyan carbon) as observed in the crystal structure.

Mutant species	Simulation ID	Superposition of the loop region between the representative structures and the binding conformation	Mutant species	Simulation ID	Superposition of the loop region between the representative structures and the binding conformation
Glu9BAla	**1**		**Lys13BAla**	**1**	
	2			**2**	
	3			**3**	
Glu10BAla	**1**		**Arg14BAla**	**1**	

	2			2	
	3			3	
Gly12BAla	1		**Glu15BAla**	1	
	2			2	
	3			3	

The RMSD values provide measures of deviation from the binding conformation in the loop region for each of the six β-Ala mutants. These in turn can give indication of the effect of different β-Ala mutations on the conformations and possible prearrangement of the loop region.

The variations in RMSD values (Table 2. 8) for the different mutant systems demonstrated that the mutations have indeed altered the conformations of the loop region. The largest RMSD deviations were found for the Gly12βAla mutant system, indicating that Gly12βAla mutation has the greatest effect on conformation of the loop region. This in turn suggests that Gly12 is important for adoption of a loop-like conformation in region Glu10-Ly13 prior to interaction with calpain. In contrast, Glu9βAla, Glu15βAla, and Lys13βAla systems showed relatively smaller deviations from the bound conformation, suggesting that these residues have a smaller influence on the loop conformation.

In order to establish if there is a trend in the effect of the mutations on the conformational preference of the loop region, the average of RMSD values for each mutant system were compared (Table 2. 8). The deviations from the binding conformation in the mutant systems are thereby ranked in the

following order: Glu15 < Arg14 < Lys13 < Glu9 < Glu10 < Gly12. The larger deviation of the loop region from the binding conformation indicates greater influence of the corresponding mutation on conformation of the loop region, and in turn indicates the greater contribution the mutated residue makes to the conformational preference of the loop region. The rank of individual residue contribution to the conformational preference of the loop region or loop prearrangement in the absence of calpain (referred to as the "conformational rank") will be in the same order as above.

Previously, Betts *et al.* have reported a rank of individual residue contribution to inhibitory activity based on the experimentally measured IC_{50} values for a series β-Ala mutants (*32*). In this "functional rank", the contributions of the mutated residues near the loop region to the inhibition activity were placed in the following order: Glu15 < Lys13 < Glu9 < Glu10 < Arg14 < Gly12.

The "conformational rank" was obtained based on only the conformational preferences of the CP1B mutants in the absence of calpain, while the "functional rank" takes into account not only the conformational factors but also other contributing factors such as the interactions between the peptide variants and calpain active site. However, comparison between the "conformational rank" from our modelling study and the "functional rank" from experimental measurements shows that the two ranks are remarkably closely correlated, with the only exception being Arg14. This difference in ranking for Arg14 may suggest that the effect of Arg14 on calpain inhibition is mainly contributed by the side chain functionality and the interactions with calpain active site, rather than its effect on the preference of the loop region conformation. However, the striking similarity of the general trend in the ranking obtained from both experimental and computational methods suggests that the differences of conformational preferences of the loop region in CP1B mutants may be a significant contributing factor to their ability to inhibit calpain.

2.4 Summary

In summary, MCSD and MD calculations have been used to demonstrate that a series of calpain inhibitors derived from natural regulator calpastatin have a preference for a adopting the loop-like conformation that is associated with binding in the absence of calpain. Inhibitory activity of these peptides can be correlated to the propensity of these peptides to prearrange in a loop-like conformation. Therefore the ability to pre-adopt loop-like conformations is an important factor for the design of effective inhibitors of calpain.

Part B:

The

3-deoxy-D-*arabino*-heptulosonate

7-phosphate synthase system

Chapter 3: Introduction to 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase

3.1 The shikimate pathway

The shikimate pathway is the seven metabolic steps beginning with the condensation of phosphoenolpyruvate (PEP) and erythrose 4-phosphate (E4P) and ending with the synthesis of chorismate (Figure 3. 1) (*43, 44*). The shikimate pathway is the biosynthetic route that is responsible for the production of essential aromatic compounds (*45*). These compounds include the aromatic amino acids, tryptophan, tyrosine and phenylalanine; folic acid, an essential cofactor for many enzymatic processes; and salicylate, used for the biosynthesis of the siderophores through which bacteria acquire iron (*46*). The pathway is found in microorganisms and plants, and has more recently been discovered in apicomplexan parasites (*47, 48*). However, the shikimate pathway is absent in higher organisms, making the enzymes of this pathway attractive as targets for the development of antimicrobial agents. Recent gene disruption studies have shown that operation of the shikimate pathway is essential for the viability of *Mycobacterium tuberculosis* (*49*), the causative agent of tuberculosis (TB), a disease that remains a significant world-wide health risk (*50*). Although effective anti-TB drugs exist, the long treatment times required, the problems of latent or persistent TB (*51*) and the proliferation of multidrug-resistant strains of *M. tuberculosis* (*52*) have all created an urgent need for the development of new antimycobacterial agents (*53*).

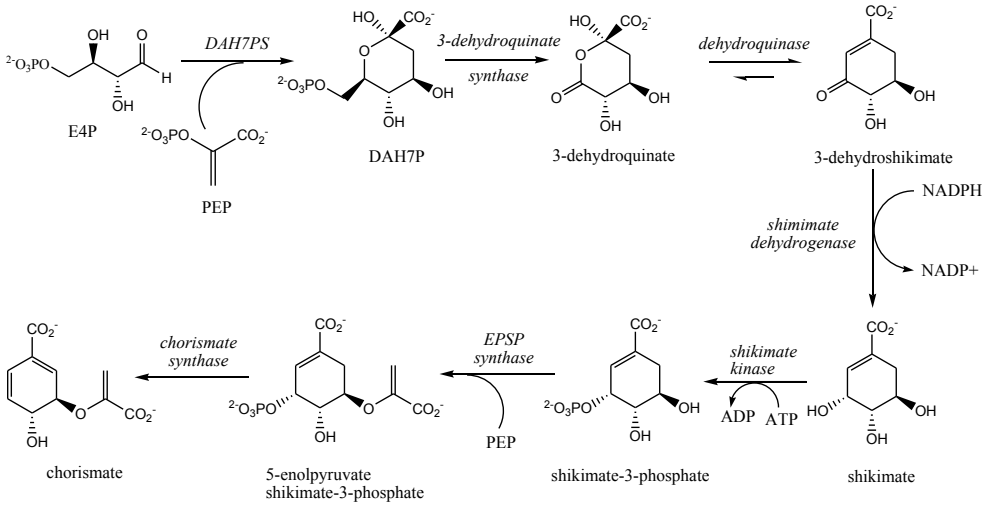


Figure 3. 1. The shikimate pathway.

3.2 3-Deoxy-D-*arabino*-heptulosonate 7-phosphate synthase

The first step of the shikimate pathway is the stereospecific aldol reaction between PEP and E4P to produce 3-deoxy-D-*arabino*-heptulosonate 7-phosphate (DAH7P) and inorganic phosphate. This reaction is catalysed by the enzyme 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase (DAH7PS). DAH7PSs are classified into two distinct types (*54*). Type I enzymes have molecular masses less than 40 kDa, and this group can be further divided into two sequence subfamilies: Iα and Iβ (*55, 56*). The type Iα and Iβ DAH7PSs have been well characterised, both structurally and functionally. The crystal structures of the type Iα enzymes from *Escherichia coli* (*57-59*) and *Saccharomyces cerevisiae* (*60, 61*), and the type Iβ enzymes from *Thermotoga maritima* (*62*) and *Pyrococcus furiosus* (*63*), reveal a common $(\beta/\alpha)_8$ TIM barrel fold (Figure 3. 2). Both the *E. coli* (Phe-sensitive) and *S. cerevisiae* (Phe- and Tyr-sensitive) enzymes have additions to their core barrels that have been shown to be associated with the binding of a single regulatory aromatic amino acid (*60, 64, 65*). Regulated type Iβ enzymes show one addition to the core barrel that is associated with allosteric regulation. This covalently attached regulatory domain appears to be either ferredoxin-like (characterised by the enzyme from *T. maritima*) (*62*) or a functional chorismate mutase fused either to the N- or C-terminus

(DAH7PS from *Bacillus subtilis* and *Porphyromonas gingivalis*) (*66*).

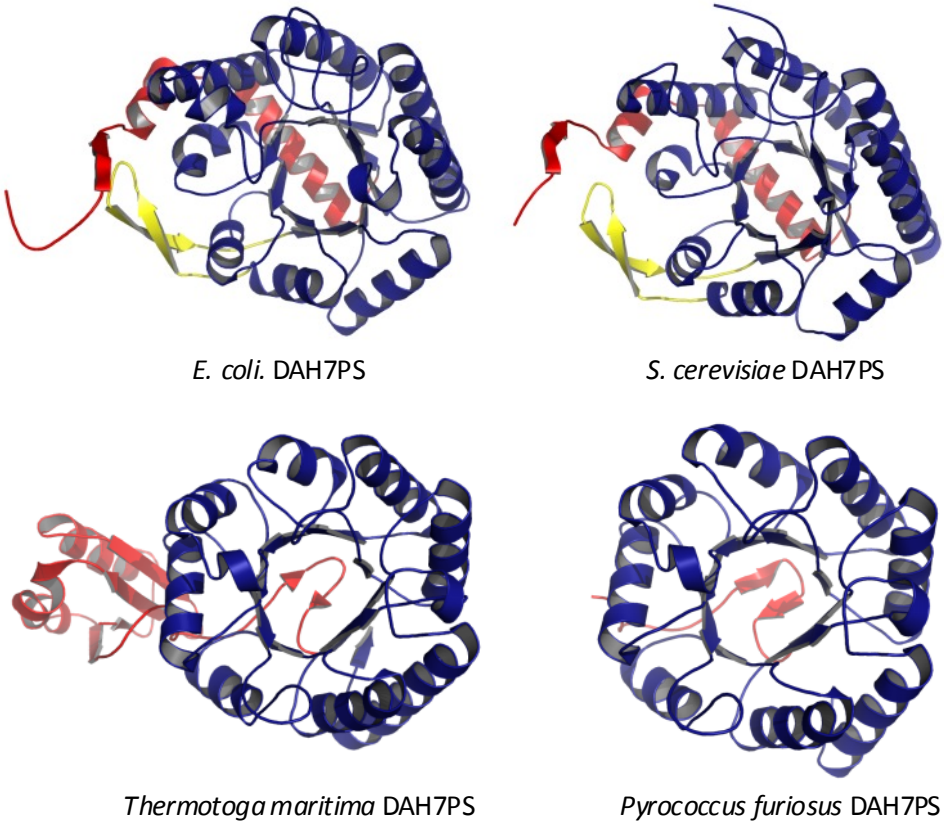


Figure 3. 2. Monomer structure of type I DAH7PS from *E. coli* (PDB 1KFL), *S. cerevisiae* (PDB 1OFB), *T. maritima* (PDB 1RZM) and *P. furiosus* (PDB 1ZCO). The additional secondary structures to the core barrel are coloured red and yellow, the main barrel structure is shown in blue.

The *M. tuberculosis* genome encodes a single DAH7PS (*Mtu*DAH7PS) that is a member of the type II family. *Mtu*DAH7PS is the only member of this type that has been structurally characterised (*67*). Type II enzymes were first identified in plants (*68, 69*), however, more recently, an increasing number of microbial examples have been identified (*54, 70*). For some organisms, type II enzymes are associated with the biosynthesis of specific secondary metabolites, in these organisms type I DAH7PS are also present (*71, 72*). However, the presence of only type II DAH7PS in a number of species including *Streptomyces* species, *Corynebacterium diphtheriae*, *Campylobacter jejuni*, *Agrobacterium tumefaciens*, *Helicobacter pylori* and several *Mycobacteria* species, supports the role of type II DAH7PS enzymes in aromatic amino acid biosynthesis. Type II enzymes are larger (>50kDa) than their type I counterparts, and the two types share very low sequence

identity (<10%). However, the core structure and catalytic machinery of *Mtu*DAH7PS is remarkably similar to those of the type I enzymes, revealing that despite their low sequence similarity these two DAH7PS types are evolutionarily related (*67*). Along with a distinct quaternary structure, *Mtu*DAH7PS displays a distinctly different decoration of the core catalytic barrel to the type I enzymes, with two additional independent subdomains, which were predicted to be linked to allosteric regulation of this enzyme.

3.3 Structural features of *Mtu*DAH7PS

The crystal structure of ligand-free wild type DAH7PS from *M. tuberculosis* has been solved previously (*67*). The monomer structure of *Mtu*DAH7PS adopts the $(\beta/\alpha)_8$ TIM barrel fold, similar to type I DAH7PS family. Two major additional structural elements were found to decorate the main barrel (Figure 3. 3A). The first addition consists of helices α0a, α0b and α0c (residues 1-80) at the N-terminus, and the second addition consists of helices α2a and α2b (residues 190-241). The two monomers in the asymmetric unit form a tight dimer, and the quaternary structure of *Mtu*DAH7PS is a homotetramer (Figure 3. 3B), which is generated by application of a crystallographic two-fold symmetry operation. The dimer interface is made up mainly by the N-terminal residues 3-10, the long helix α2, and the α0b-α0c loop from the N-terminal addition, with a solvent accessible surface area of 1885 $Å^2$. The interface between the two dimers to form the tetramer is less extensive with solvent accessible surface area of 954 $Å^2$, and is made up by the additional α2b helix and α2b-β3 loop, with minor contributions from the core α1 helix.

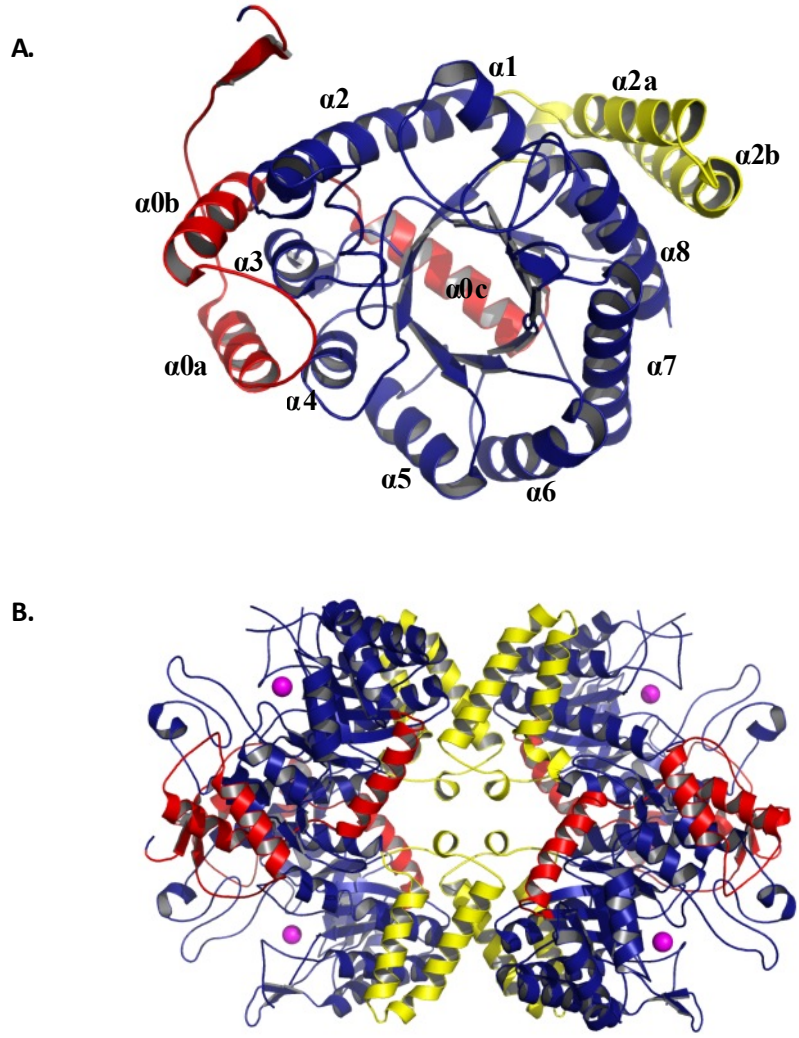


Figure 3. 3. A. Monomer and B. tetramer structure of ligand-free wild type *Mtu*DAH7PS (PDB 2B7O). The main barrel structure is shown in blue, the additional helices at the N-terminus are coloured red and the additional helices α2a and α2b are coloured in yellow. The bound metal ions (Mn^{2+}) are displayed in (B) as magenta spheres to indicate the position of the active site.

The active site of *Mtu*DAH7PS is located at the C-terminal end of the barrel, where it is clearly indicated by the presence of the bound metal ion, Mn^{2+} and PEP (Figure 3. 4). The active site residues are contributed by the C-terminal ends of the β-strands and α-helices. The metal binding site is formed by Cys87, His369, Glu411 and Asp441.

The PEP binding site is defined by a network of hydrogen bonds between the protein and the PEP phosphate and carboxylate groups. The phosphate group of PEP forms hydrogen bonds with Glu283, Arg284, Lys306, Arg337 and two water molecules. The carboxylate group of PEP forms hydrogen bonds with Arg126, Lys306 and one additional water molecule. One sulfate ion is

also observed to bind in the active site. This sulfate ion interacts with Arg135, Arg284 and Ser136 and is likely to indicate the position of the E4P phosphate group upon binding of the substrate.

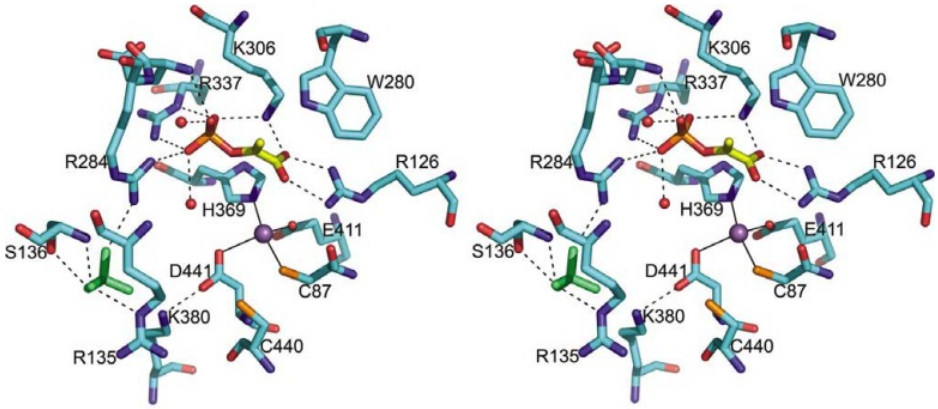


Figure 3. 4. Stereo view of the active site of *Mtu*DAH7PS. The active site metal ion (Mn^{2+}) is shown as magenta sphere, the substrate PEP is shown with yellow carbon atoms, and the sulfate ion bound in the active site is shown in green. The active site residues are displayed with blue carbon atoms.

3.4 Feedback inhibition studies with *Mtu*DAH7PS

Some initial studies on *Mtu*DAH7PS have been conducted by Celia J Webby (*73*). The three aromatic amino acids (Trp, Phe and Tyr) and chorismate were tested, individually and in combination, as potential feedback regulators of *Mtu*DAH7PS. Results showed that single amino acids do not inhibit the activity of *Mtu*DAHP7S, but combinations of Trp+Phe, and Trp+Tyr reduce the activity of the enzyme to 54% and 78% respectively. This observation indicated that *Mtu*DAH7PS is inhibited synergistically with combination of aromatic amino acids.

3.5 Outline of Part B of this thesis

Part B of this thesis explores the structure and function of *Mtu*DAH7PS using various techniques of molecular modelling. Firstly, protein crystal structure refinements were conducted and those crystal structures of

*Mtu*DAH7PS in complex with various ligand molecules are described in Chapter 4. Secondly, the reaction mechanism and roles of active site residues were investigated in Chapter 5, through docking calculations (both rigid docking and induced fit docking) of a series of designed active site inhibitors. Finally, Chapter 6 discusses the molecular basis of the communication mechanism of allosteric regulation in *Mtu*DAH7PS.

Chapter 4: Synergistic allostery and series of crystal structures of wild type and mutant *Mtu*DAH7PS in complex with different ligands

4.1 Introduction

4.1.1 Synergistic feedback inhibition of Mtu*DAH7PS by combinations of amino acids*

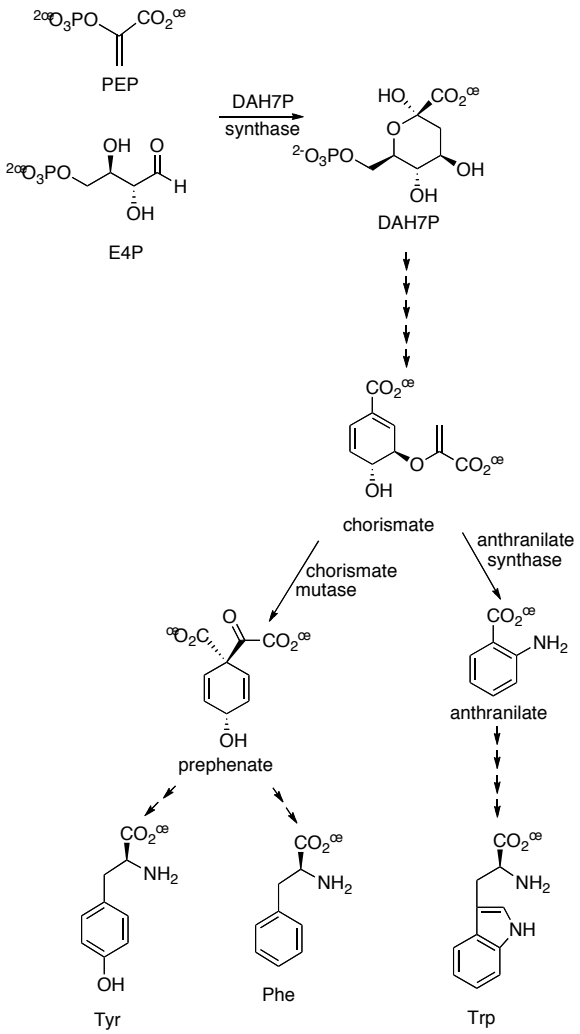


Figure 4. 1. DAH7PS reaction and shikimate pathway

As the first enzyme, DAH7PS is a major control point for shikimate pathway flux (Figure 4. 1). Several organisms express two or more isozymes of this

enzyme that show different sensitivity to the pathway end products. For example, *E. coli* and *Neurospora crassa* each produce three isozymes, with each enzyme individually inhibited by Phe, Tyr or Trp (*74, 75*). For *S. cerevisiae* there are two differentially regulated DAH7PS sensitive to either Phe or Tyr (*60*).

M. tuberculosis does not produce any isozyme of DAH7PS, and *Mtu*DAH7PS is regulated with a more sophisticated mechanism. Kinetic experiments have shown that *Mtu*DAH7PS is not inhibited by any single aromatic amino acid (Phe, Tyr or Trp alone), but that combinations of aromatic amino acids (Trp and Phe, and to a lesser extend Trp and Tyr) give significant enzyme inhibition (*76*). No inhibition was observed in the presence of both Tyr and Phe, a combination that has recently been reported to inhibit the activity of the *M. tuberculosis* chorismate mutase in complex with *Mtu*DAH7PS (*77*).

Steady-state kinetic experiments showed that PEP binding is unaffected by the presence of the inhibitory combinations of aromatic amino acids (Trp+Phe and Trp+Tyr), however, a dramatic change in cooperativity and a significant decrease in affinity was observed for E4P binding (*76*).

4.1.2 Crystallography and structural refinement

Macromolecular X-ray crystallography has become one of the most critical experimental techniques in studies of enzymes and other biological macromolecules. X-ray crystal structures not only show quaternary and tertiary structures of the protein fold, but can also be used to elucidate details of the enzyme active site and ligand binding sites at atomic level, and provide valuable insights to the structural-function relationships of the enzyme. The crystal structures can be used to guide further studies by either experimental or computational means. Structures can be used to inform design of drug-like compounds at designated sites on the enzyme and in other cases they are usually used as the starting point in computational calculations, such as molecular dynamics (MD) simulations and QM/MM

calculations.

Once appropriate forms of crystals are formed in the laboratory, the structure of an enzyme can be determined by passing X-ray beams through the crystal and record the scattering patterns of the X-ray caused by the electrons in the enzyme molecule. Because the structures in each unit cells are repeated throughout the entire volume of a crystal in a periodic fashion, it can be treated as a 3D diffraction grating. The scattering of the X-ray beams is governed by the geometry of the unit cell and the wavelength of the X-ray beam; and the intensities of the diffracted ray are governed by the atoms that constitute the structures in the unit cell. Therefore the entire structure of the enzyme is encoded in the diffracted X-ray pattern. The diffracted beams are also called "reflections", and the recorded intensities of the diffracted X-ray contain only the amplitude information, but not the phases of the reflections. Therefore, a Fourier transformation of the structure factors (*F* values) is calculated that contains both the amplitude and phase information of the observed reflections. This usually leads to an initial approximate electron density distribution in the crystal, which can be improved in an iterative fashion, eventually converging at an appropriate structural model of the protein (*78*).

Up to this point, the data generated (the electron density map) can still be considered as experimental results. However, to obtain a 3D model of the enzyme structure, the electron density map has to be interpreted by adding individual atoms and molecules to the model. The model is then refined through a number of rounds to achieve the best possible agreement between the atomic model and the electron density map. The process of structural refinement can therefore be considered as a "modelling" technique. Because the atomic interpretation can be subjective, especially when modelling those ambiguous regions of the electron density map, slightly different atomic models of the enzyme structure can be obtained with the same electron density map.

In order to investigate the mechanism of the observed synergistic inhibition of *Mtu*DAH7PS by combinations of aromatic amino acids (Trp+Phe and Trp+Tyr), crystal structure information is required to identify binding sites and binding modes of the ligand molecules, and any structural changes that may be caused by ligand binding. In this chapter of the thesis, the technique of crystal structure refinement was employed and the physical meanings of the final 3D atomic models were discussed.

Prior to the start of the work in this chapter, crystal structures of *Mtu*DAH7PS soaked with Trp+Phe, Trp-only and Phe-only had been solved and initially refined by Celia J Webby (*73*). As part of the studies described in this thesis, further refinements for these crystal structures were conducted, and refinements for a series of other structures of *Mtu*DAH7PS were carried out to enable a fuller understanding of the binding of allosteric ligands to *Mtu*DAH7PS.

The refinement method used for the crystal structures described in this chapter is the same as described in the paper by Webby *et al*. (*76*).

4.2 Crystal structures of wild type *Mtu*DAH7PS

4.2.1 Crystal structures of Mtu*DAH7PS in complex with inhibitory combinations of aromatic amino acids*

4.2.1.1 Mtu*DAH7PS in complex with Trp+Phe*

The crystal structure of *Mtu*DAH7PS in complex with the inhibitory amino acid combination Trp+Phe was obtained firstly by soaking Trp and Phe at high ligand concentration (2.2 mM Trp and 2.2 mM Phe) (refined at 2.0 Å), and then by co-crystallisation at low ligand concentration (100 μM Trp and Phe) (refined at 1.95 Å) (Table 4. 1).

Table 4. 1. Refinement statistics for MtuDAH7PS in complex with Trp and Phe obtained by soaking and co-crystallisation.

Structures		**Trp+Phe soaked**	**Trp+Phe cocrystal**
Space group		$P3_221$	$P3_221$
Unit cell dimensions (Å)	***a=b***	207.56	204.54
	c	207.56	66.46
Refinement resolution range (Å)		32.7-2.0	179.8-1.95
Number of reflections used in refinement		105110	113806
***R* factor**		0.163	0.153
R_{free}		0.191	0.168
Number of non-hydrogen atoms		6976	7800
Mean *B* value (Å^2)		25.42	33.9
rms deviations from ideality	**Bond lengths (Å)**	0.016	0.012
	Bond angles (deg.)	1.962	1.318
Residues in the most favored region of Ramachandran plot (%)		97.6	96.8
PDB code		3KGF	3RZI

Trp+Phe soaked structure. The final model of Trp+Phe soaked structure contains three Phe and two Trp molecules. The asymmetric unit contains two molecules with the general $(\beta/\alpha)_8$ TIM barrel fold. Two major additions decorate the basic barrel structure in each monomer. The first addition (residues 1-76) is the extended N-terminus and the three additional helices (α0a, α0b and α0c) (Figure 4. 2a). The second addition consists of a pair of helices (α2a and α2b) and a coil (residues 193-239) that extend the α2-β3 loop (Figure 4. 2a).

The two molecules in the asymmetric unit form a tight dimer, which further associates by application of crystallographic two-fold symmetry to form an *Mtu*DAH7PS homotetramer (Figure 4. 2). The dimer interface has a solvent-accessible surface area of 1858 Å^2, and is formed mainly by elements from the N-terminal extension to the core $(\beta/\alpha)_8$ barrel, specifically residues 3-10, which form a two-stranded beta sheet across the interface, and the α0b-α0c loop, together with the long helix α2. The interface between the two dimers, through which the tetramer is generated, is less extensive, with solvent accessible surface area of 1065 Å^2, and is formed by extra-barrel

elements, helix α2b and the loop α2b-β3, with minor contributions from the core α1 helix. The ligands Phe and Trp bind at the dimer and tetramer interfaces respectively, forming hydrogen bonds as well as hydrophobic contacts with the extra-barrel elements. One additional Phe per dimer is located between the α0a, α0b and α3 helices and the α-β4 loop of one of the monomer units.

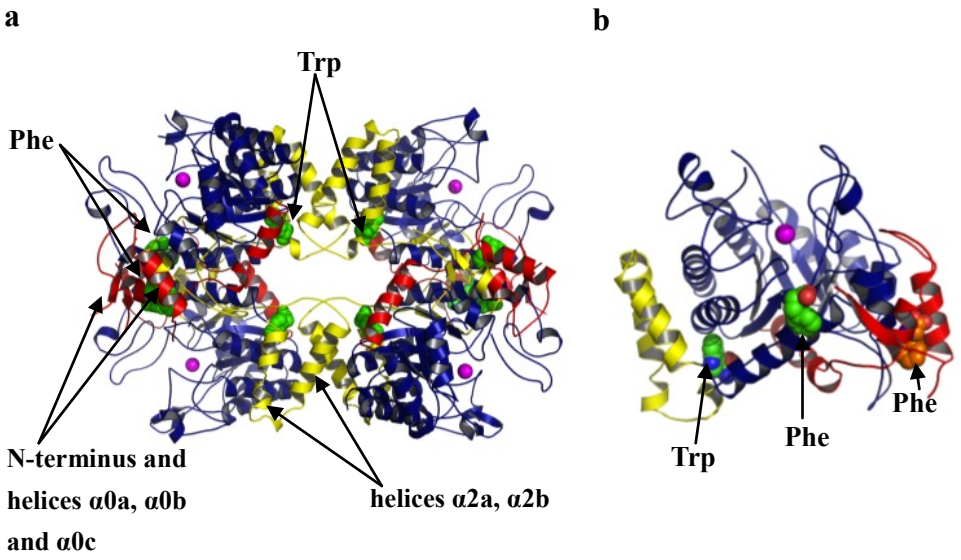


Figure 4. 2. a. Homotetramer and **b.** monomer structure of *Mtu*DAH7PS in complex with Phe and Trp obtained by soaking (3KGF). Ligand molecules are displayed in green. In both figures, the N-terminal extra-barrel elements (the N-termini and α0a-α0c helices) are shown in red, and the additional α2a, α2b helices are shown in yellow. Active site metal ions are shown as purple spheres and the ligands are in sphere form. The Phe binding in secondary site is shown with orange carbon atoms, and the Phe in primary site and the Trp ligand are shown with green carbons.

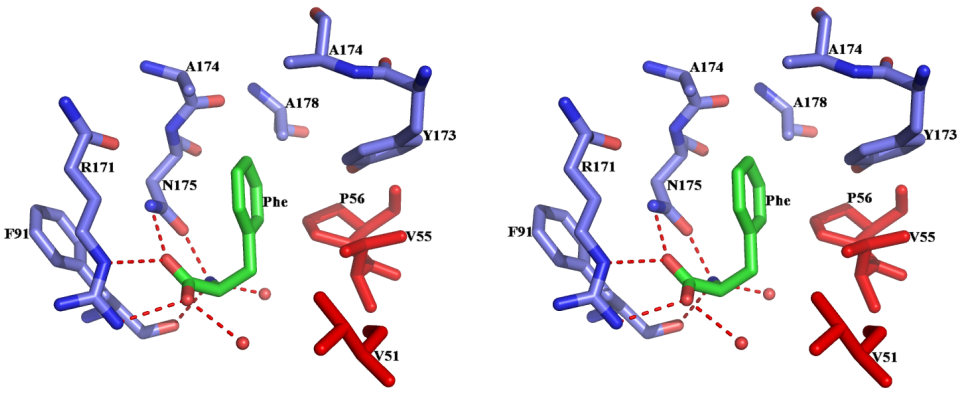


Figure 4. 3. Stereo view of the primary Phe binding sites in the tight dimer interface in Trp+Phe soaked structure (3KGF). Residues 3-10 are excluded for clarity.

The primary Phe binding sites in the Trp+Phe soaked structure (3KGF) are

located at the interface of the tight dimer, with binding involving interactions with residues from both monomers (Figure 4. 3). There is one primary Phe binding site in each subunit. The α-carboxylate group of each Phe ligand forms hydrogen bonds with the side chains of Arg171 and Asn175 and the α-amino group interacts with the side chain of Asn175 and the main-chain carbonyl oxygen of Phe91. The hydrophobic ring of Phe is buried in a pocket formed by residues from both subunits including: Ala174, Ala178, Pro56*, Val55*, Ala174*, and Tyr173* (where * indicates residues from the other subunit of the tight dimer). These binding pockets are "capped" by the extended N-termini of both monomers, residues 3-10, with Trp3 from subunit B positioned between the two Phe ligands, and contributing to the binding pockets for both Phe molecules (Figure 4. 4).

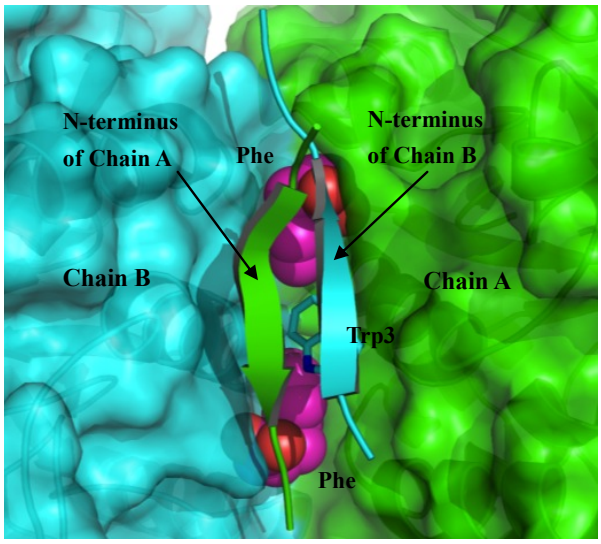


Figure 4. 4. The "cap" of the primary Phe binding sites. Trp3 from subunit B is positioned between the two Phe molecules bound at the dimer interface. The N-termini in both subunits are displayed in ribbon form, and the rest of the enzyme is displayed by surface.

A secondary Phe binding site is observable in one monomer in each of the tight dimers of the tetrameric complex in this Trp+Phe soaked crystal structure (3KGF) (Figure 4. 5). The binding pocket for this site is formed by residues from the additional α0a and α0b helices, the α3 helix and the α3-β4 loop. The α-carboxylate group of the Phe ligand bound here interacts with the side chains of Arg23 and Arg256. The α-amino group is hydrogen bonded to the backbone carbonyl oxygen of Arg256 and the side chain of Glu253. The hydrophobic ring sits in a pocket formed by nearby Leu

residues 26, 18, 271, 261 and 259. The helices α0a and α0b, which make up the binding pockets of the secondary Phe binding sites, are slightly closer to each other in subunit A and there is insufficient space for a Phe ligand to bind in this subunit. In subunit B, in contrast, the residues on the α0a helix (residues 16-27) have shifted in position so that the bound Phe can be accommodated.

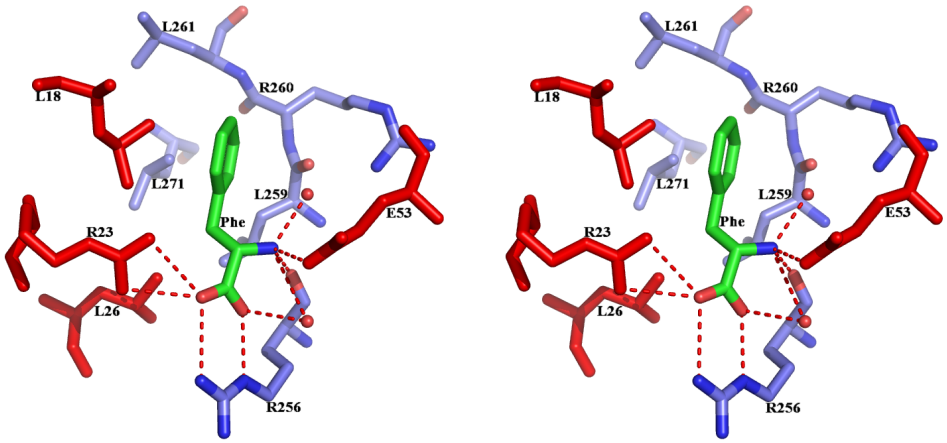


Figure 4. 5. Stereo view of the secondary Phe binding site in subunit B of Trp+Phe soaked structure (3KGF).

The Trp binding sites are located at the interface between the two tight dimers that make up the tetramer (Figure 4. 2). Each molecule of Trp is bound in a cavity between the extra-barrel pair of helices (α2a and α2b) and the core barrel helix α1 in each monomer. The Trp binding site is predominately formed by residues of the α2b and α1 helices, and the α2b-β3 loop, which is also involved in tetramer association. In subunit A (Figure 4. 6), the α-carboxylate group of Trp forms a salt bridge with the Lys123 side chain. The α-amino group interacts with the backbone carbonyl oxygen atoms of Asn237 and Thr240. The N^{ε} on the Trp indole ring is hydrogen bonded to the backbone carbonyl oxygen of Ala192. The hydrophobic ring sits in a relatively deep binding pocket formed by hydrophobic residues, Val111, Leu107, Leu194, Ala241 and Val197. Similar interactions are also observed for Trp bound in subunit B.

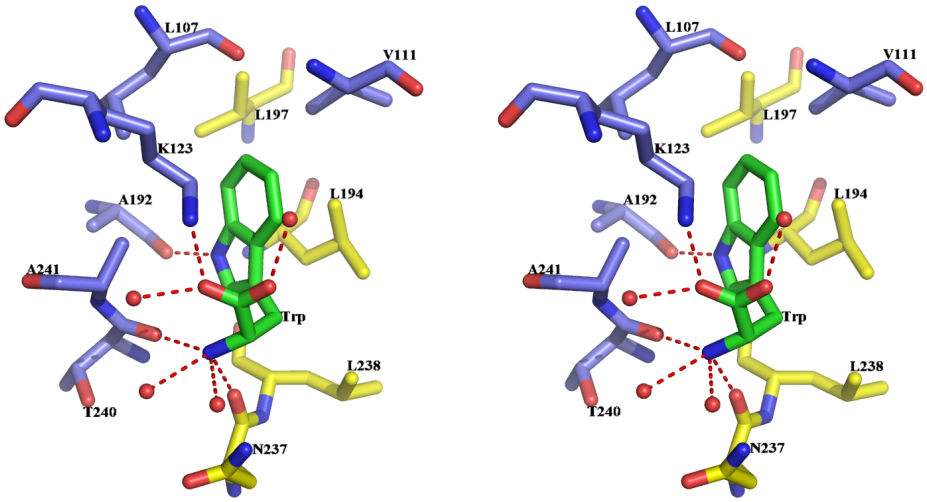


Figure 4. 6. Stereo view of the Trp binding sites in Trp+Phe soaked structure (3KGF).

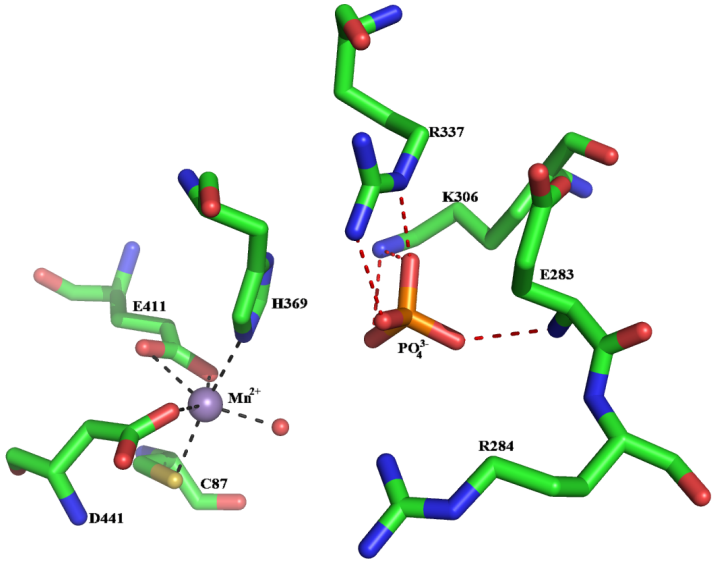


Figure 4. 7. The metal binding site and the phosphate ion bound in the active site of the Trp+Phe soaked structure (3KGF). The active site residues are displayed with green carbons, the metal ion is shown as a sphere.

The active site structure of the Trp+Phe soaked *Mtu*DAH7PS is very similar to that of the ligand-free structure previously reported (PDB 2B7O), with one metal ion Mn^{2+} bound in the active site of each monomer. The metal binding site is formed by four protein ligands, Cys87, His369, Asp441, and both of the carboxylate oxygen atoms of Glu411. These residues fill up five positions of a pseudo-octahedral metal coordination site. In subunit A (Figure 4. 7), a water molecule (HOH278) occupies the sixth coordination site, whereas in subunit B, the sixth ligand is absent, but the geometry of the metal binding site is the same as that in subunit A.

One phosphate ion is bound in the active site of each monomer (Figure 4. 7), where it interacts with the side chains of the nearby Lys306, Arg337, the backbone NH group of Glu283, and four water molecules. The same interactions are observed in both chains. This phosphate ion binds in the same position as the phosphate group of PEP when it is bound in the active site of ligand-free *Mtu*DAH7PS structure (PDB 2B7O) (Figure 4. 8).

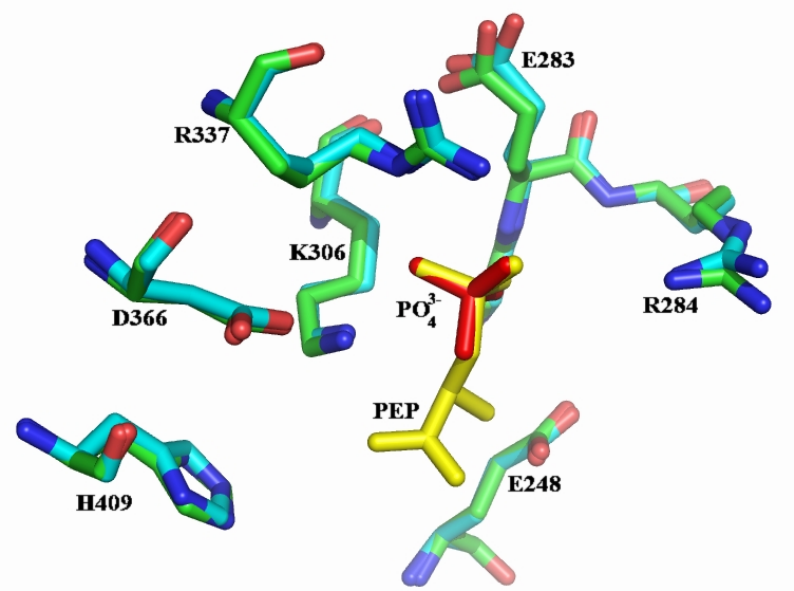


Figure 4. 8. The active site residues from Trp+Phe soaked structure (3KGF) superimposed onto the ligand-free structure determined previously (2B7O). The PEP molecule bound in the ligand-free structure (2B7O) is shown in yellow and the phosphate ion found in the Trp+Phe soaked structure is coloured in red.

Trp+Phe co-crystal structure. Since the Trp+Phe soaked crystal structure discussed above showed two distinct Phe binding sites, it is not clear which binding site of Phe is critical for inhibition. Therefore, the structure of *Mtu*DAH7PS in complex with both Phe and Trp was determined by co-crystallisation at concentrations of inhibitors known to result in substantial loss of enzymatic activity (100 μM of each ligand). This co-crystallised Trp+Phe bound structure (Figure 4. 9) closely resembles the Trp+Phe soaked structure (root-mean-square distance of 0.644 Å, matching 1753 C^{α} atoms). The solvent accessible surface area of the tight dimer interface is 1875 $Å^2$ and that of the tetramer interface is 1024 $Å^2$.

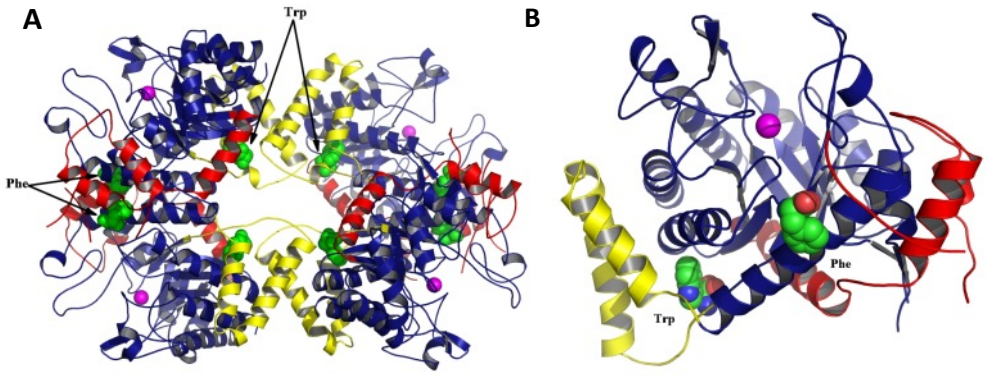


Figure 4. 9. A. Homotetramer and B. Monomer of *Mtu*DAH7PS co-crystallised with Trp and Phe ligands (3RZI).

In the structure determined by co-crystallisation, two Phe molecules bind in each of the tight dimer interfaces and two Trp molecules are located in each of the tetramer interfaces (Figure 4. 9). Thus one tetrameric biological unit binds four Trp and four Phe molecules simultaneously. The interactions between the bound ligand molecules and their binding sites at the interfaces are consistent with those observed in the soaked crystal structure.

In the Trp+Phe soaked structure (3KGF), an additional Phe binding was observed and this site was occupied in one subunit of the asymmetric unit. However, in the co-crystallised structure (3RZI), the secondary Phe binding site is unoccupied, and its absence in the co-crystallised structure indicates that this site is occupied only at the high Phe concentrations (2 mM) used in the soaking experiment. This observation confirms that the binding sites at the two interfaces as observed in the co-crystal structure are responsible for the synergistic inhibition we have observed for the Trp+Phe combination, and the occupation of the secondary Phe binding site is not essential for inhibition.

4.2.1.2 Crystal structure of MtuDAH7PS in complex with Trp and Tyr

The crystal structure of *Mtu*DAH7PS in complex with the other inhibitory combination of aromatic amino acids (Trp+Tyr) was determined and refined at 2.8 Å). This crystal was also obtained by soaking (

Table 4. **2**).

Table 4. 2. Refinement statistics for *Mtu*DAH7PS in complex with Trp and Tyr obtained by soaking.

Structure		**Tyr+Trp**
Space group		$P3_221$
Unit cell dimensions (Å)	***a=b***	202.777
	c	66.747
Refinement resolution range (Å)		175.67-2.76 (2.83-2.76)
Number of reflections used in refinement (test set)		37979 (2242)
***R* factor**		0.183
R_{free}		0.251
Number of non-hydrogen atoms		7251
Mean *B* value (Å^2)		32.747
rms deviations from ideality	**Bond lengths (Å)**	0.025
	Bond angles (deg.)	3.673
Residues in the most favored region of Ramachandran plot (%)		89.88
PDB code		Not deposited

The final model of Trp+Tyr soaked structure contains three Tyr and two Trp molecules. This structure is very similar to the Trp+Phe soaked crystal structure (3KGF) (rms distance of 0.441 Å, matching 806 C^{α} atoms). The solvent accessible surface area of the tight dimer interface is 1761 Å^2 and that of the tetramer interface is 1022 Å^2.

In the Tyr+Trp soaked structure, Tyr molecules are found to bind in the primary Phe binding site that is observed in the Trp+Phe soaked structure (Figure 4. 10). The interactions between the Tyr ligand and the residues in the binding site are the same as those formed when Phe is bound, with the α-carboxylate group forming hydrogen bonds with the side chains of Arg171 and Asn175 and the α-amino group interacting with the side chain of Asn175 and the main-chain carbonyl oxygen of Phe91. No additional interactions are formed between the enzyme residues and the Tyr hydroxyl group.

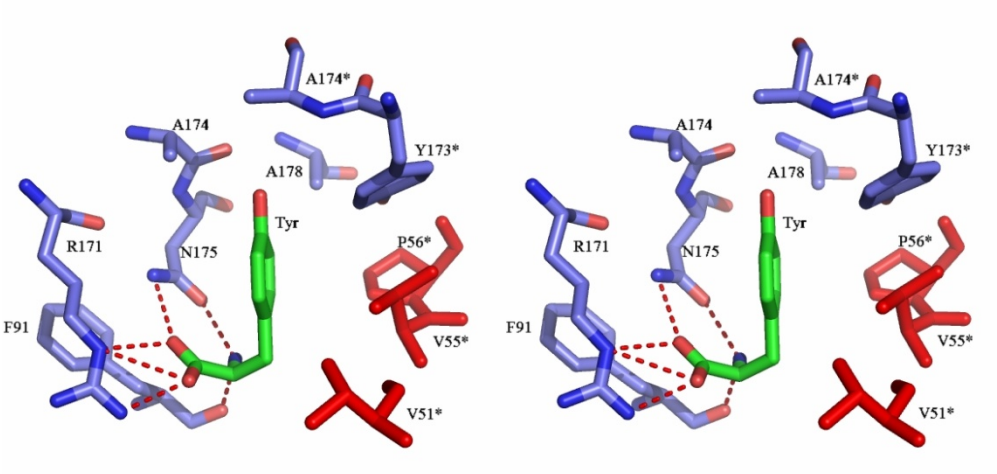


Figure 4. 10. Tyr molecule binding in the primary Phe binding sites in the tight dimer interface. Residues 3-10 are excluded for clarity.

Similar to Trp+Phe soaked structure (3KGF), one Tyr molecule is found to bind in the secondary Phe binding site (Figure 4. 11). The same interactions between Phe molecule and the residues in the secondary Phe binding site are also observed when Tyr is bound in this site. In addition, Tyr molecule forms one extra hydrogen bond between the hydroxyl group and the backbone carbonyl of Pro16. This suggests that the so-called secondary Phe binding site may actually have higher affinity for Tyr molecules over Phe molecules.

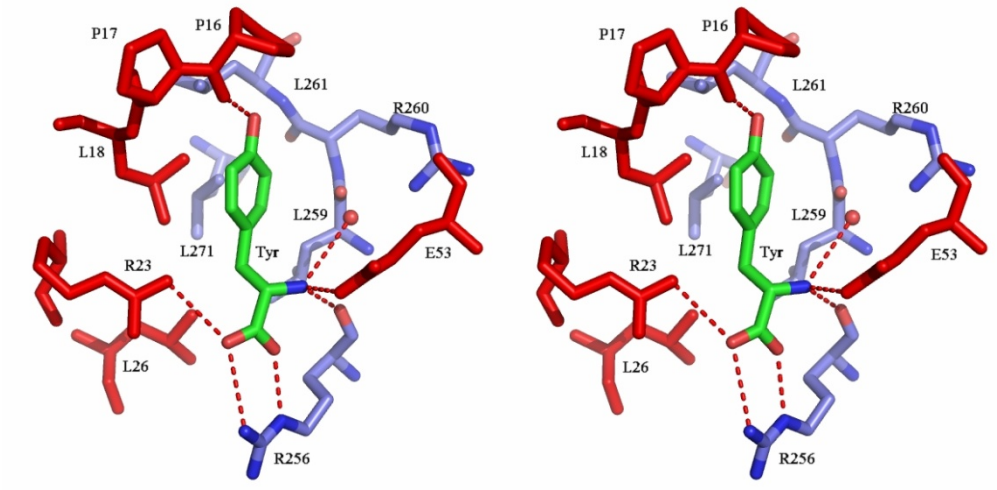


Figure 4. 11. Tyr molecule binding in the secondary Phe binding sites.

Trp molecules still bind in the same binding sites at the tetramer interface in the Trp+Tyr soaked structure, and the interactions between Trp molecules and the residues in the binding sites are the same as observed in Trp+Phe

soaked structure (3KGF).

4.2.2 Crystal structure of ligand-free Mtu*DAH7PS (Thesit-free)*

Previously the ligand-free wild type crystal structure of *Mtu*DAH7PS was determined (PDB 2B7O) (*67*). In this crystal structure, a detergent Thesit molecule that was present in the crystallisation condition can be observed to bind in the tight dimer interface. After the preliminary refinement of the Trp+Phe soaked structure (3KGF) conducted by Celia J Webby, it was found that in the presence of Trp and Phe ligand molecules, the detergent Thesit molecule was displaced by two Phe molecules binding in the same dimer interface (Figure 4. 12). Therefore, to enable direct comparisons of ligand-free and various ligand-bound crystal structures, the ligand-free structure of *Mtu*DAH7PS that is also free of bound Thesit molecules was determined and refined at 2.3 Å (Table 4. 3).

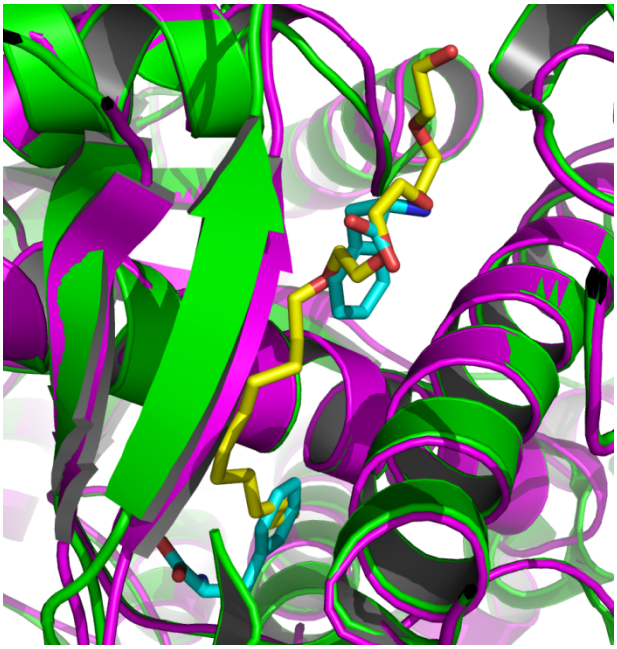


Figure 4. 12. The Thesit molecule (yellow) in the original ligand-free structure 2B7O (green) is displaced by two Phe molecules (cyan) in the Trp+Phe soaked structure 3KGF (magenta).

In the final model of the Thesit-free structure, both molecules in the asymmetric unit contain the full polypeptide chain (Residues 1 to 462). The final model contains one PEP molecule, one phosphate ion, two manganese

ions, three sulfate ions, one chloride ion, and six glycerol molecules. The solvent accessible surface area of the tight dimer is 1885 $Å^2$ and that of the tetramer interface is 954 $Å^2$.

The new ligand-free Thesit-free structure of *Mtu*DAH7PS (3NV8) is essentially identical to the original Thesit-bound ligand-free structure (2B7O), with the α-carbons in both chains in the two structures match with RMSD of 0.3 Å. The major differences only occur in structurally disordered regions including the α3-β4 loop (residues 265-268), β7-α7 loop (residues 375-381), and α2b-β3 loop (residues 234-241). In the tight dimer interface where the Thesit molecule was bound in the original crystal structure (2B7O), only slight changes are observed for the N-terminal residues 10 to 16 in Chain B. All other residues within 5 Å of the Thesit molecule still adopt the same conformation in the Thesit-free structure.

Table 4. 3. Refinement statistics for ligand-free *Mtu*DAH7PS (Thesit-free).

Structure		**Thesit-free**
Space group		$P3_221$
Unit cell dimensions (Å)	***a=b***	203.54
	c	66.42
Refinement resolution range (Å)		39.37-2.25 (2.31-2.25)
Number of reflections used in refinement (test set)		69535 (3648)
***R* factor**		0.180
R_{free}		0.208
Number of non-hydrogen atoms		7098
Mean *B* value ($Å^2$)		36.9
rms deviations from ideality	**Bond lengths (Å)**	0.008
	Bond angles (deg.)	2.157
Residues in the most favored region of Ramachandran plot (%)		95.1
PDB code		3NV8

4.2.3 Comparison of Trp+Phe soaked structure with Theist-free structure

Since the Trp+Phe cocrystal structure (3RZI) and Trp+Tyr soaked structure are both very similar to the Trp+Phe soaked structure (3KGF), only one of these is used to compare with ligand-free structure (3NV8) to inform any major structural changes that occur upon simultaneous binding of the inhibitory combinations of aromatic amino acids to *Mtu*DAH7PS. The crystal structure of the Trp+Phe soaked *Mtu*DAH7PS complex (3KGF) was superimposed on to the new Thesit-free ligand-free structure (3NV8), in order to identify any structural differences caused by ligand binding and to identify a possible mechanism for inhibition. Ligand binding was found to cause no significant change in the overall structure of the *Mtu*DAH7PS enzyme (Figure 4. 13).

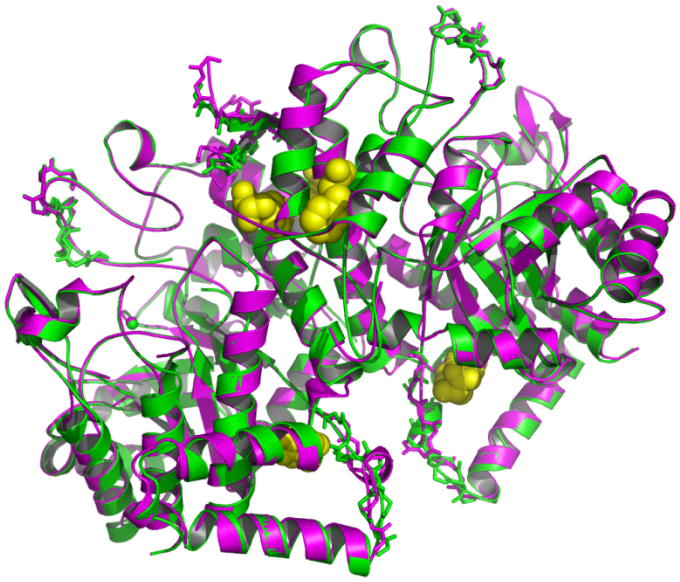


Figure 4. 13. Superposition of the Thesit-free structure 3NV8 (magenta) and the Trp+Phe soaked structure 3KGF (green) of *Mtu*DAH7PS. The Phe and Trp ligands are in yellow sphere form. The backbones of residues in regions with minor conformational changes are displayed in stick form.

The two structures match with an RMSD in C^{α} atomic positions of only 0.7 Å over the entire dimer. The greatest displacements were in the Phe and Trp binding sites, the N-termini, and the disordered α3-β4 loop (residues 263-268). Upon binding of the Trp and Phe ligands, residues from β4 to α8 in both monomers (residues 273-462) show very little change in C^{α} positions,

indicating these secondary structures are not affected significantly.

Binding of both Trp and Phe ligands causes the extended β2-α2 loop (residues 135-144) in the active site, which contains the highly conserved KPRS motif for type II DAH7PS enzymes, to become more disordered. These residues have been shown to be associated with the binding of substrate E4P in other DAH7PS enzymes (*61, 62*). This change is more clearly observed in subunit A, in which the loop shows alternative positions for residues 135-139 (Figure 4. 14). This change in the active site loop is not observed in the Trp-only and Phe-only structures.

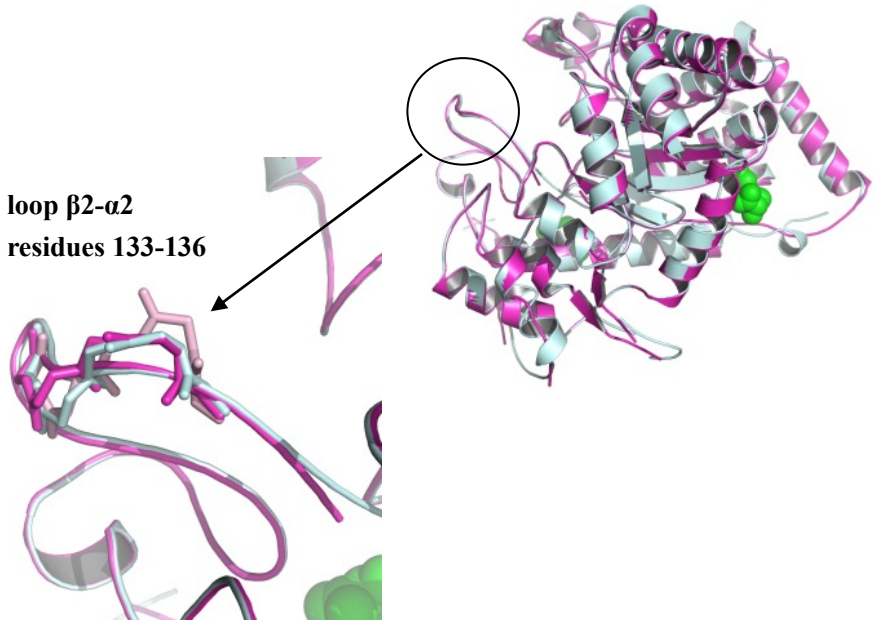


Figure 4. 14. Superposition of chain A and zoom in view of the active site loop containing the KPRS motif in the Trp+Phe soaked structure (3KGF, shown in magenta), and the ligand-free Thesit-free structure (3NV8, shown in cyan) of *Mtu*DAH7PS. The backbone of residues 133-136 is displayed in stick form, and the alternative conformation found in Trp+Phe soaked structure is shown in pink.

In the tight dimer interface, the binding of Trp and Phe simultaneously results in a subtle movement of residues of the α2 helix. This allows Ser189 on the α2 helix to form new interactions with Ser62 on the α0c helix of the other monomer (Figure 4. 15). Another interaction gained at the dimer interface upon ligand binding is the hydrogen bond between Arg236 and Asn237 of the other monomer, with Asn237 also contributing to the tetramer interface (Figure 4. 16). In spite of Phe binding into the tight dimer interface,

the backbone C^{α} atom positions are not affected for most amino acids near the Phe binding sites except for those on the N-termini. For both monomers, the N-terminal residues (1-10) have shifted in position (Table 4. 4), with displacements up to 6.2 Å in chain B and 3.1 Å in chain A. The displacement is greater in chain B because in the ligand-bound structure Trp3 sits in between the two bound Phe molecules, near where Val5 was found in the ligand-free structure.

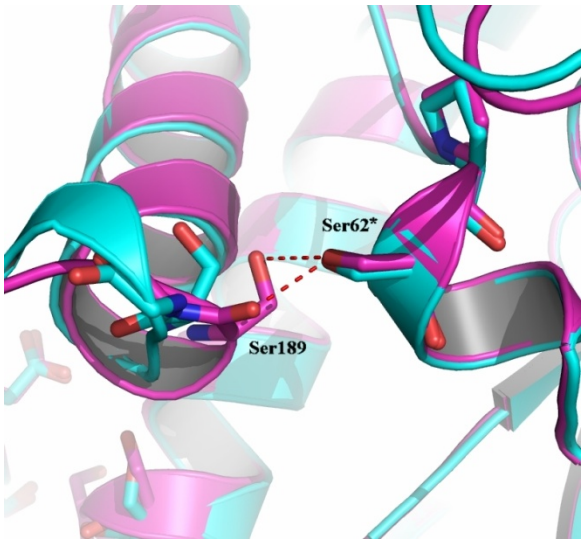


Figure 4. 15. The flip of peptide bond at Ser189 upon binding of Trp and Phe. Ligand-free structure (3NV8) is shown in cyan and Trp+Phe soaked structure is shown in magenta (3KGF).

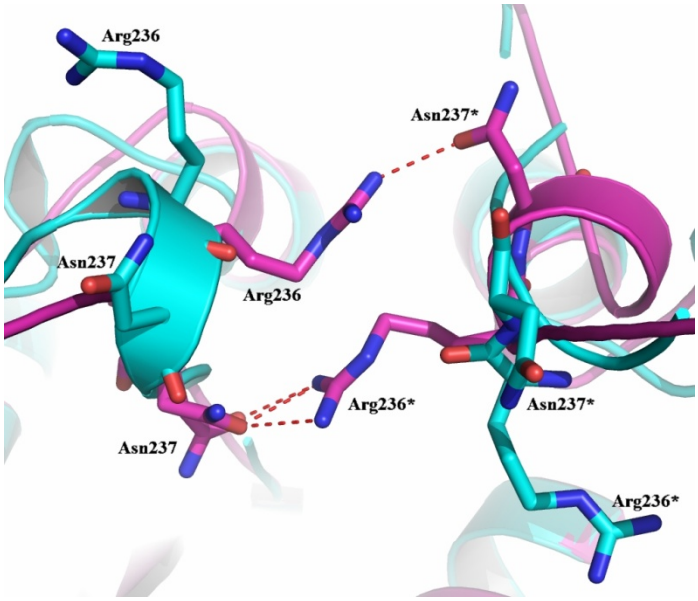


Figure 4. 16. The new interaction at dimer interface between Arg236 and Asn237 upon Trp and Phe binding. Ligand-free structure (3NV8) is shown in cyan and Trp+Phe soaked structure is shown in magenta (3KGF).

Table 4. 4. Displacement of C^{α} atoms upon ligand binding. Only residues with C^{α} displacement greater than 1 Å in either subunit are listed.

Residue	C^{α} Displacements (Å) upon Trp and Phe Binding	
	Subunit A	Subunit B
Met1		4.97
Asn2		4.96
Trp3	1.50	6.19
Thr4	1.52	5.37
Val5	1.33	5.71
Asp6	1.35	6.13
Ile7	1.51	5.56
Pro8	1.42	6.14
Ile9	2.10	4.80
Asp10	3.07	
Leu15		2.88
Pro16	0.40	2.20
Asp138	1.59	0.56
Ser189	1.60	1.84
Gly190	1.46	1.75
Leu191	1.09	1.30
Ser193	1.00	1.10
Leu196	0.87	1.09
Asp199	1.00	1.14
Arg202	0.83	1.20
Arg206	1.05	1.23
Ser208	0.85	1.12
Pro209	0.89	1.03
Gly211	1.05	1.03
Met228	0.89	1.12
Ser229	1.06	1.21
Ala230	1.02	0.99
Gly232	1.04	0.94
Val233	2.56	2.19
Ala234	5.38	4.89
Asp235	3.35	3.36
Arg236	1.32	1.27
Asn237	4.47	4.53
Leu238	4.07	4.20
Gln239	6.84	6.98
Thr240	3.18	2.62
Gly264	0.34	1.18
Asp265		1.99
Asp266		1.87
Gly267		1.45

Upon binding of Phe in the Trp+Phe soaked structure, the side chain of Arg171 in chain B becomes disordered and shows one alternative conformation that interacts with the carboxylate group the Phe molecule. This displaces the side chain of Asp10 in chain A which was previously occupying the same site in the ligand-free structure, and results in the loss of the salt bridge between Arg171 of chain B and the Asp10 of chain A (Figure 4. 17). Due to the shifts of N-terminal residues in subunit B, Arg171 in chain A gained a salt bridge with Asp6 and does not show any alternative conformation.

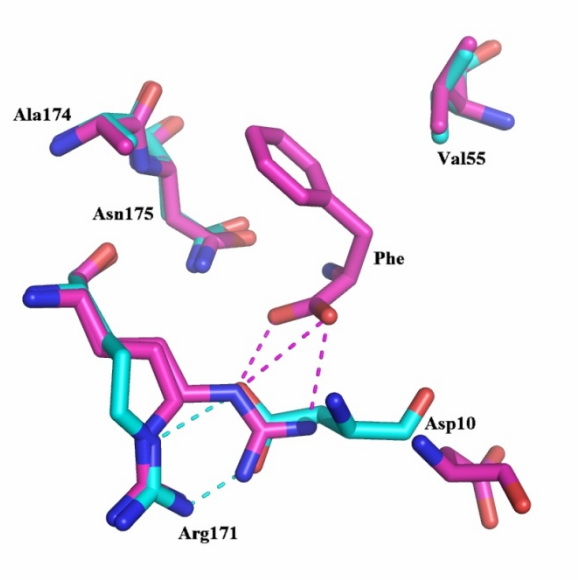


Figure 4. 17. Alternative conformation of Arg171 in chain B upon Phe binding in the dimer interface. Ligand-free structure (3NV8) is shown in cyan and Trp+Phe soaked structure is shown in magenta (3KGF).

The residues contributing to the tetramer interface remain the same upon ligand binding although they undergo considerable structural changes. Considerable structural changes of the tetramer interface occur in the α2b-β3 loop, while the α2a and α2b helices display slight movements near the Trp binding sites. Upon binding of the Trp ligands, residues 232-241 (in the α2b-β3 loop) show the most significant displacements in C^{α} atoms (Table 4. 4). As a result the backbone carbonyl groups of Gln237 and Thr240 both interact with the amino group of Trp molecule. The side chain conformations of residues Leu194 and Val111 are changed in order to form the hydrophobic

pocket for Trp binding. The side chains of Lys123 in both monomers also moved slightly to interact with the carboxylate groups of the bound Trp molecules (Figure 4. 18).

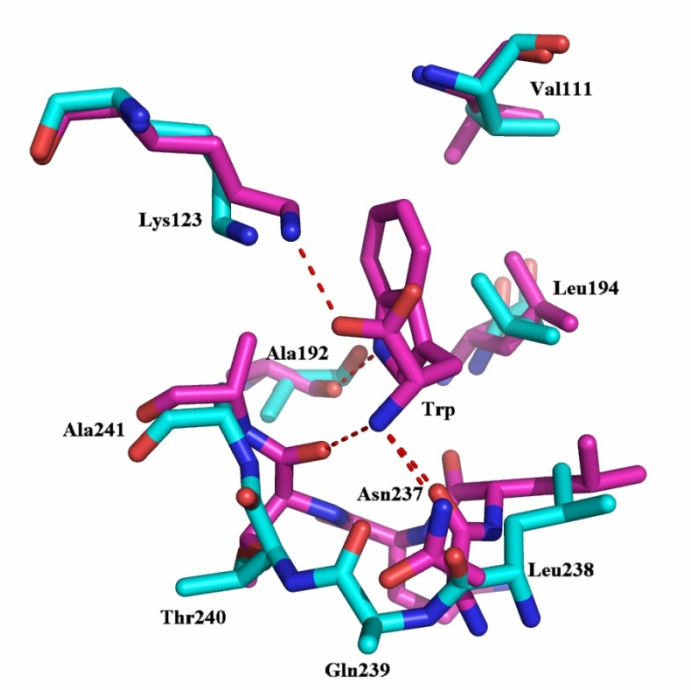


Figure 4. 18. Changes in residue conformations around tetramer interface upon binding of Trp.

4.2.4 Crystal structure of Mtu*DAH7PS in complex with single amino acids*

4.2.4.1 Phe-only soaked structure

The crystal structure of *Mtu*DAH7PS in complex with Phe alone obtained by soaking was refined at 3.0 Å (

Table 4. **5**). The final model contains five phosphate ions and five Phe molecules in the asymmetric unit. Three of the Phe molecules occupy the same primary and secondary binding sites as in the Trp+Phe soaked structure (3KGF), with two bound in the tight dimer interface and one bound in subunit B near helices α0a, α0b and α3. The remaining two Phe molecules are observed to bind in the Trp binding sites in the tetramer interface. The solvent accessible surface area of the tight dimer is 1746 $Å^2$ and that of the tetramer interface is 1030 $Å^2$.

Table 4. 5. Refinement statistics for *Mtu*DAH7PS in complex with Phe alone obtained by soaking.

Structure		**Phe-only**
Space group		$P3_221$
Unit cell dimensions (Å)	***a=b***	204.65
	c	66.61
Refinement resolution range (Å)		39.6-3.0 (3.08-3.0)
Number of reflections used in refinement (test set)		29566 (1574)
***R* factor**		0.212
R_{free}		0.269
Number of non-hydrogen atoms		6756
Mean *B* value (Å^2)		32.7
rms deviations from ideality	**Bond lengths (Å)**	0.012
	Bond angles (deg.)	1.160
Residues in the most favored region of Ramachandran plot (%)		92.1
PDB code		3NUD

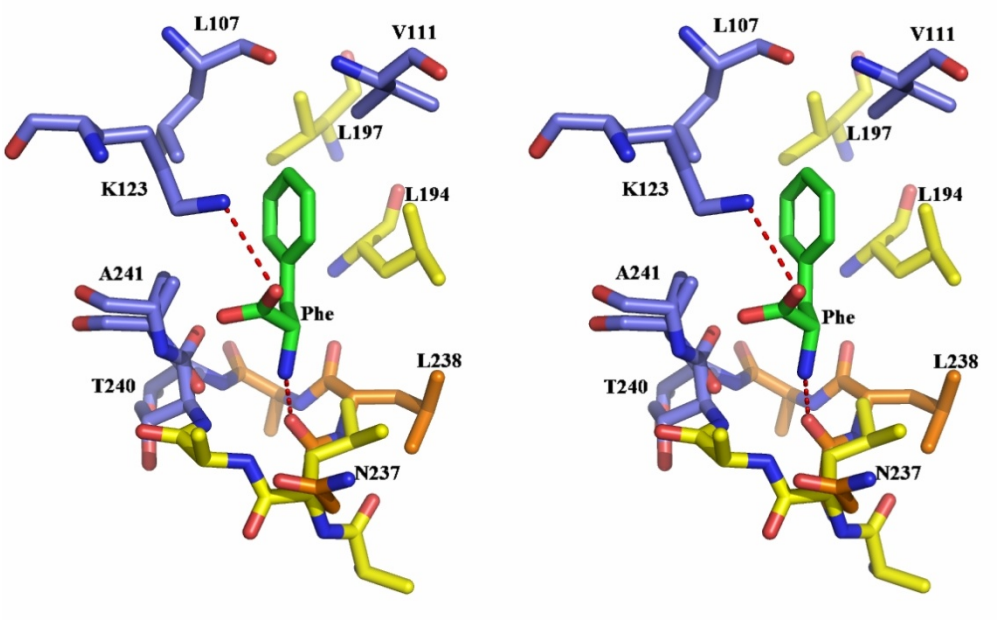


Figure 4. 19. Stereo views of Phe in the Trp site at the tetramer interface in the Phe-only structure (3NUD). The alternative positions for residues 237-239 are displayed with orange carbons.

In the Phe-only structure (3NUD) the primary and secondary Phe binding sites are essentially the same as in the Trp+Phe soaked structure (3KGF), with minor changes in side chain conformations of Arg171, and small shifts

in the positions of the N-terminal residues. The interactions between the bound Phe ligands and the enzyme are retained.

Interestingly, in the Phe-only structure obtained by soaking in high concentration of Phe ligand (2 mM), Phe molecules are observed to also occupy the Trp binding sites at the tetramer interface and show similar interactions (Figure 4. 19). The α-carboxylate group of the Phe interacts with the side chain of Lys123, and the hydrophobic ring sits in the same pocket as that of the Trp ligand when it is bound. Unlike Trp, the α-amino groups of the bound Phe ligands only show interaction with the backbone carbonyl oxygen of Asn237, but not Thr240. The loop comprising residues 237-242 is more disordered than in the Trp+Phe soaked structure, and shows at half occupancy an alternative position similar to that observed in the ligand-free structure (3NV8).

4.2.4.2 Tyr-only soaked structure

Table 4. 6. Refinement statistics for *Mtu*DAH7PS in complex with only Tyr obtained by soaking.

Structure		**Tyr-only**
Space group		$P3_221$
Unit cell dimensions (Å)	***a=b***	202.777
	c	66.747
Refinement resolution range (Å)		175.67-2.69 (2.76-2.69)
Number of reflections used in refinement (test set)		41510 (2959)
***R* factor**		0.156
R_{free}		0.190
Number of non-hydrogen atoms		7168
Mean *B* value (Å^2)		47.9
rms deviations from ideality	**Bond lengths (Å)**	0.028
	Bond angles (deg.)	2.349
Residues in the most favored region of Ramachandran plot (%)		92.08
PDB code		Not deposited

The structure of *Mtu*DAH7PS in complex with only Tyr was obtained by soaking in high concentration of Tyr ligand (2 Mm) and refined at 2.7 Å (Table 4. **6**). The final structure contains three Tyr ligands in the asymmetric

unit. The solvent accessible surface area of the tight dimer is 1736 Å^2 and that of the tetramer is 973 Å^2.

The Tyr molecules in the Tyr-only structure are observed to bind in the primary and secondary Phe sites in the same way as observed in the Tyr+Trp soaked structure.

With the absence of Trp molecules, Tyr molecules are not found in the Trp binding sites as was observed for Phe molecules in the Phe-only structure (3NUD), despite Tyr being similar in structure with Phe. This may suggest that the Trp binding pocket does not efficiently accommodate the additional hydroxyl group of Tyr molecule.

4.2.4.3 Trp-only structure

The crystal structures of *Mtu*DAH7PS in complex with only Trp molecules were obtained by soaking in 2mM Trp (refined at 2.5 Å) and by co-crystallisation at very high Trp concentrations (1 mM) (refined at 2.3 Å) (Table 4. 7).

Table 4. 7. Refinement statistics for MtuDAH7PS in complex with Trp alone obtained by soaking and co-crystallisation.

Structures		**Trp-only soaked**	**Trp-only cocrystal**
Space group		$P3_221$	$P3_221$
Unit cell dimensions (Å)	***a=b***	204.68	204.71
	c	204.68	66.26
Refinement resolution range (Å)		38.7-2.5 (2.57-2.5)	44.24-2.25
Number of reflections used in refinement (test set)		52324 (2815)	72258
***R* factor**		0.189	0.143
R_{free}		0.228	0.169
Number of non-hydrogen atoms		6871	8150
Mean *B* value (Å^2)		28.2	23.97
rms deviations from ideality	**Bond lengths (Å)**	0.016	0.018
	Bond angles (deg.)	1.485	1.67
Residues in the most favored region of Ramachandran plot (%)		95.6	97.2
PDB code		3NUE	Not deposited

Trp-only soaked structure. The final model of Trp-only soaked structure (3NUE) contains two phosphate ions, two manganese ions, four sulfate ions, one chloride ion, one Thesit molecule and two Trp molecules. The Trp binding sites in the Trp-only structure are essentially identical to that in the Trp+Phe soaked structure (3KGF), with Trp molecules forming the same interactions with the enzyme. The solvent accessible surface area of the tight dimer interface is 2125 $Å^2$ and that of the tetramer interface is 1047 $Å^2$.

One Thesit molecule is bound in the Phe binding sites in the tight dimer interface. The N-terminal residues has shifted in position in comparison with the Trp+Phe soaked structure, however the positions of these residues are essentially identical to those found in the previously reported ligand-free wild type structure, which also has a Thesit molecule bound at the same site (PDB 2B7O). This suggests that the shift in N-terminal residues only occurs upon Phe binding in the tight dimer interface. This shift in position may be caused by the positioning of Trp3 of subunit B between the two bound Phe molecules.

Trp-only co-crystallised structure. The crystal structure of *Mtu*DAH7PS in complex with Trp alone was determined by co-crystallisation at high ligand concentration (1 mM of Trp). This co-crystallised Trp-only bound structure closely resembles the Trp+Phe co-crystal and soaked structures determined above, with rms distance of 0.270 Å matching 849 C^{α} atoms, and rms distance of 0.149 Å matching 867 C^{α} atoms, respectively. The solvent accessible surface area of the tight dimer interface is 2034 $Å^2$ and that of the tetramer interface is 1189 $Å^2$.

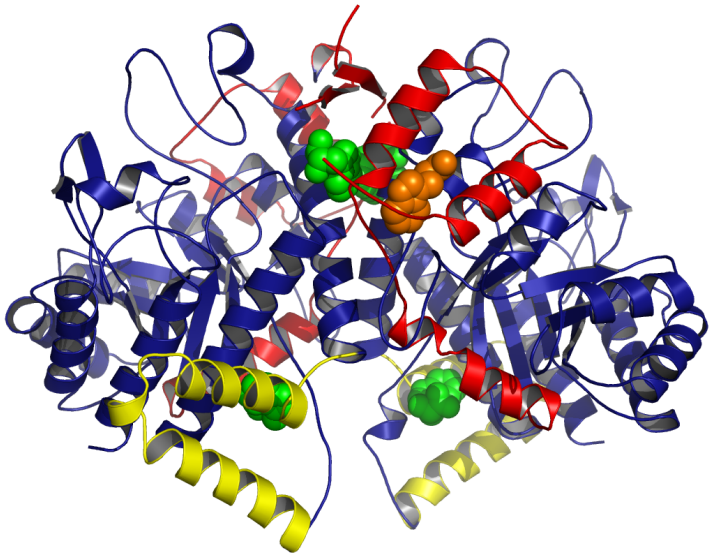


Figure 4. 20. The dimer in the asymmetric unit of *Mtu*DAH7PS co-crystallised with only Trp at high ligand concentration. The basic barrel structure is shown in blue; the N-terminal additions are coloured red, and the additional helices at tetramer interfaces are coloured yellow. The Trp molecules bound in the interfaces are shown in green, and the Trp molecule bound in the secondary Phe site is coloured in orange.

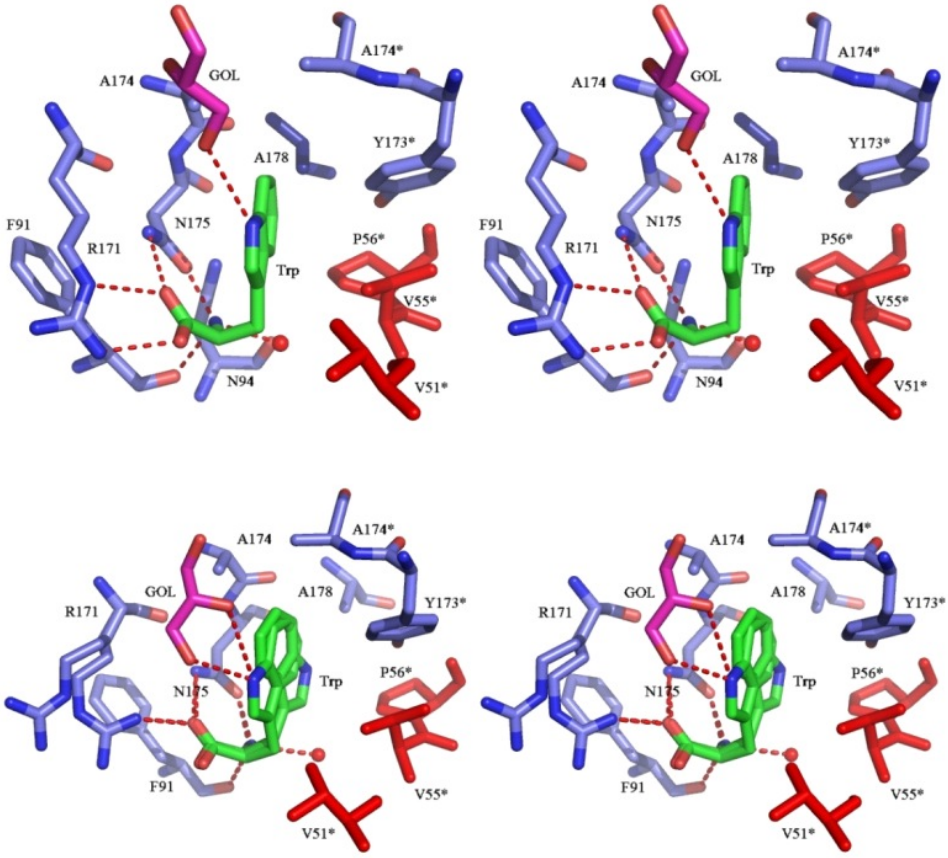


Figure 4. 21. Trp molecule bound in primary Phe site. Trp molecule binding in chain A is shown on top, and Trp molecule binding in chain B is shown on the bottom, both of the alternative conformations of Trp molecules binding in chain B are displayed. The glycerol molecule positioned in between the two Trp molecules is shown in magenta.

There are five Trp molecules in the asymmetric unit (Figure 4. 20), two Trp molecules bind in the normal Trp site with the same binding mode as observed in all other crystal structures with Trp bound, and surprisingly, two Trp molecules are found to bind in the primary Phe binding sites, and one Trp molecule is found in the secondary Phe binding site. In other words, Trp molecule is observed to bind in all possible ligand binding sites in this co-crystal structure. This suggests that these ligand binding sites can be accommodative enough to fit Trp molecule at very high concentration.

The carboxylate and amino groups of Trp molecules bound in the primary Phe site form the same hydrogen bond interactions with the enzyme residues as when Phe is bound (Figure 4. 21). The Trp molecule bound in chain A forms an additional hydrogen bond with a glycerol molecule that is positioned between the two Trp molecules in the dimer interface, and this glycerol molecule also forms an additional hydrogen bond with the Trp molecule bound in chain B (Figure 4. 21). This glycerol molecule is in the same position as the Trp3 residue of chain B in the Trp+Phe soaked and co-crystal structures. As a result, the N-terminal shift upon Phe binding is not observed in the Trp-only co-crystal structure. The N-terminal residues (residues 1-10) in the Trp-only co-crystal structure adopts the same conformations and positions as in the ligand-free wild type structure (Thesit-free). The Trp molecule bound in chain B is disordered and two alternative conformations can be modelled into the electron density map. This suggests that although the primary Phe binding sites at the dimer interface can accommodate the bigger Trp molecules, they are not as tight binding as Phe molecules, and only occupy the primary Phe binding sites at such high ligand concentration (1 mM).

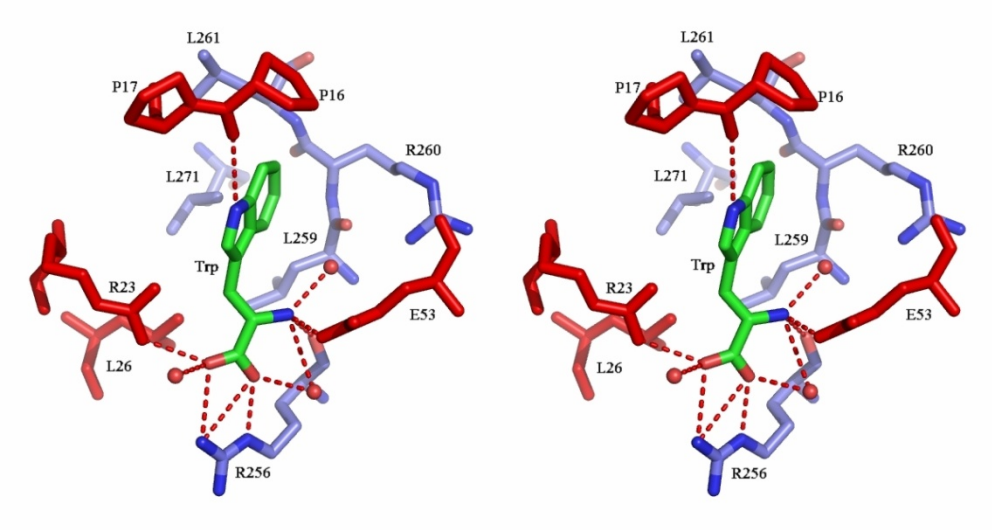


Figure 4. 22. Trp molecule bound in the secondary Phe site.

In the Trp-only co-crystal structure, one Trp molecule is observed to bind in the secondary Phe binding site with the same binding mode as when Phe is bound (Figure 4. 22). The interactions between the enzyme residues and the backbone carboxylate and amino groups of the Trp molecule are the same as those observed in the Phe-only and Trp+Phe soaked structure. Interestingly, Trp molecule is also capable of forming one extra hydrogen bond with the backbone carbonyl group of Pro16, in a manner similar to when Tyr is bound in this site in the Tyr-only and Tyr+Trp structure.

4.3 Implications on molecular mechanism of feedback allosteric regulation of *Mtu*DAH7PS

Seven crystal structures of *Mtu*DAH7PS in complex with various ligands (Trp+Phe soaked, Trp-only soaked, Phe-only soaked, Tyr+Trp soaked, Tyr-only soaked, Trp-only cocrystal, Trp+Phe cocrystal) was determined and one ligand-free wild type structure without the Thesit molecule bound was also obtained. Comparison of these wild type structures shows clearly that there is no gross conformational change within *Mtu*DAH7PS associated with any ligand binding event (

Table 4. **8**).

Table 4. 8. RMSD values between various ligand-bound crystal structures of *Mtu*DAH7PS and the ligand-free structure (3NV8).

Structures	*rms distance compare to Thesit-free ligand-free structure (Å)*	*number of matching C^{α} atoms*
Trp+Phe soaked	0.334	808
Trp-only soaked	0.354	798
Phe-only soaked	0.328	778
Tyr+Trp soaked	0.652	846
Tyr-only soaked	0.255	841
Trp-only cocrystallised	0.355	823
Trp+Phe cocrystallised	0.489	806

The synergistic nature of allosteric regulation of *Mtu*DAH7PS by only combinations of amino acids is quite unusual. The fact that there is no gross conformational change associated with ligand binding is even more intriguing. Experimental kinetic measurements have shown that the binding of both Trp and Phe affects the binding of the substrate E4P by reducing the enzyme's affinity for this substrate. From comparison of crystal structures, it can be observed that the loop responsible for E4P binding (loop β2-α2) shows an alternative conformation in the Trp+Phe soaked structure (3KGF), suggesting increased flexibility in this region may be associated with ligand binding. These observations suggest that mechanism of allosteric regulation and the communication among distant binding sites of *Mtu*DAH7PS may arise from an alteration of protein dynamics upon ligand binding. Therefore, MD simulations can be employed to study the change in dynamics within the enzyme structure in detail, and may help to elucidate the molecular basis of this unusual synergistic allostery in *Mtu*DAH7PS.

4.4 Crystal structures of mutant *Mtu*DAH7PS

Crystal structures of two mutants of *Mtu*DAH7PS were obtained (
Table 4. **9**). Both mutations (R23A, R256A) are located in the secondary Phe binding site, and both of the mutated Arg residues participate in interactions with the carboxylate group of the bound ligand at this site (Phe, Tyr or Trp).

Table 4. 9. Refinement statistics for mutant crystal structures of *Mtu*DAH7PS.

Structures		**R23A**	**R256A**
Space group		$P3_221$	$P3_221$
Unit cell dimensions	***a* (Å)**	203.479	203.77
	***b* (Å)**	203.479	203.77
	***c* (Å)**	66.34	66.22
Resolution range (Å)		40.37-2.41	62.23-2.37
Number of reflections		112702	60903
***R* factor**		0.164	0.168
R_{free}		0.183	0.199
Number of non-hydrogen atoms		7643	7211
Average B factors (Å^2)		19.12	23.63
Rms deviations from ideality	**Bond lengths (Å)**	0.014	0.006
	Bond angles (deg.)	2.849	1.502
Residues in most favored region of Ramachandran plot (%)		95.23	96.10
PDB code		Not deposited	Not deposited

4.4.1 R23A

The crystal structure of *Mtu*DAH7PS R23A mutant was refined at 2.41 Å to a final *R* factor of 0.164 (R_{free}=0.183). This mutant structure is very similar to the ligand-free Thesit-free wild type crystal structure (3NV8), with RMSD of 0.253 Å matching 870 C^{α} atoms (Figure 4. 23). The solvent accessible surface area of the tight dimer is 1918 Å^2 and that of the tetramer interface is 967 Å^2.

The overall conformation of *Mtu*DAH7PS was not altered by the R23A mutation, apart from the residues on loop α2-α2a and α2b-β3 (residues 188-195 and 233-244) which showed a slightly different conformation to that of the ligand-free wild type structure, indicating the flexible nature of these loops. Inspection of the mutant structure at position 23 indicated clearly that the mutation was successful (Figure 4. 24). In chain B of the ligand-free wild type structure (3NV8), one sulfate ion is observed to interact with the side chain of Arg23. However in the mutant structure, the sulfate ion is not observed to bind in this site due to loss of the side chain functionality of

Arg23 (Figure 4. 24).

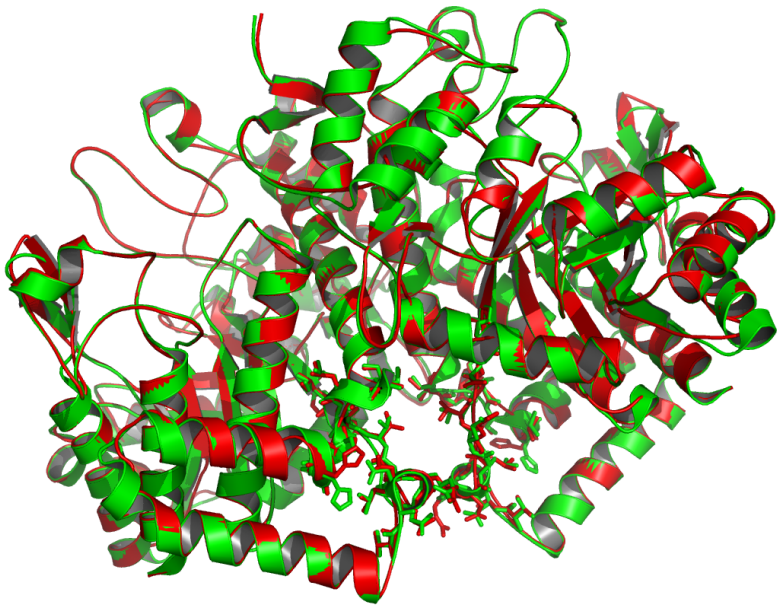


Figure 4. 23. Superposition of crystal structure of R23A mutant (red) with ligand-free Thesit-free wild type structure 3NV8 (green). The positions with slightly different conformations are highlighted by displaying the residues in stick form (residues 188-195, 233-244).

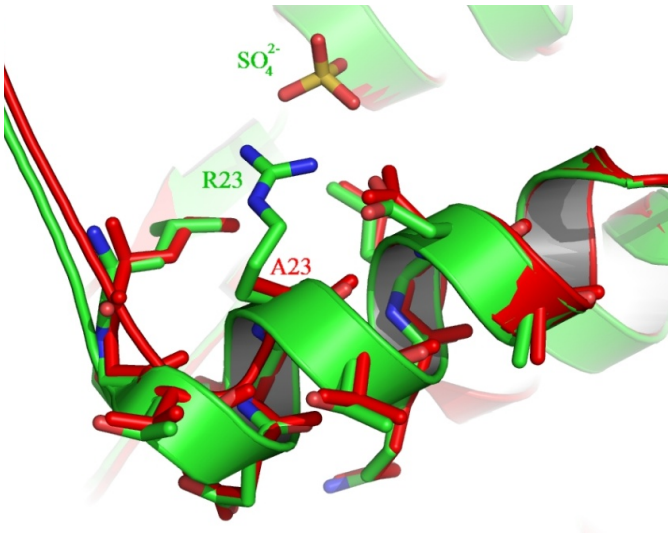


Figure 4. 24. Superposition of R23A mutant crystal structure (red) with the ligand-free wild type crystal structure (green) at the mutation site (R23).

4.4.2 R256A

The crystal structure of R256A mutant of *Mtu*DAH7PS was refined at 2.37 Å to a final *R* factor of 0.168 (R_{free}=0.199). This mutant structure is also very close to the ligand-free wild type crystal structure (3NV8), with rms distance of 0.228 Å matching 873 C^{α} atoms (Figure 4. 25). The solvent accessible surface area of the tight dimer interface is 1878 $Å^2$ and that of the tetramer interface is 975 $Å^2$.

Similar to R23A mutant, the mutation site of R256A crystal structure (Figure 4. 26) has lost the ability to bind the sulfate ion that is observed in the wild type structure.

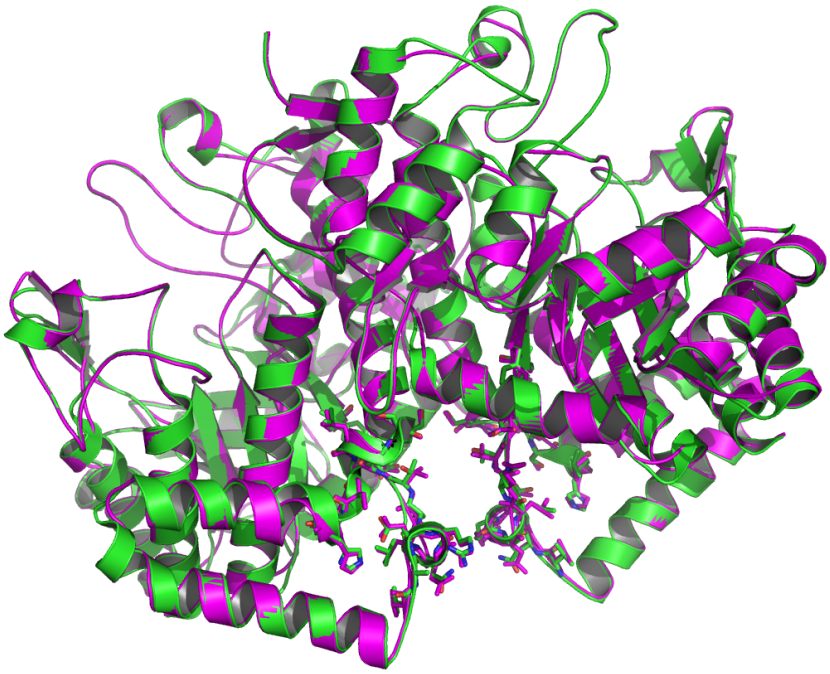


Figure 4. 25. Superposition of crystal structure of R256A mutant (magenta) with ligand-free wild type structure 3NV8 (green). The positions with slightly different conformations are highlighted by displaying the residues in stick form (residues 188-195, 233-244).

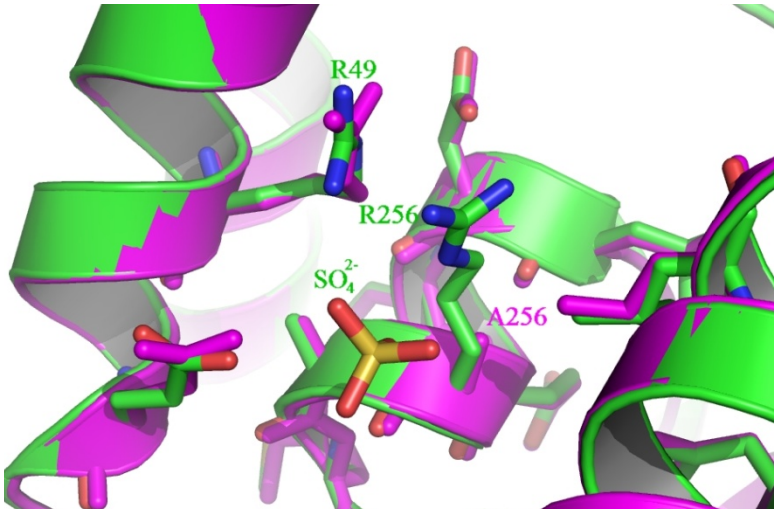


Figure 4. 26. Superposition of R256A mutant crystal structure (magenta) with the ligand-free wild type crystal structure 3NV8 (green) at the mutation site (R256).

4.5 Summary

*Mtu*DAH7PS is synergistically inhibited by combinations of Trp+Phe and to a lesser extent by Trp+Tyr, but not by any single aromatic amino acids. In order to understand this unique mechanism of allosteric regulation of *Mtu*DAH7PS, crystal structures of wild type *Mtu*DAHP7S in complex with different aromatic amino acids and two mutants of *Mtu*DAH7PS were refined and examined in this chapter.

Crystal structures of *Mtu*DAH7PS in complex with Trp+Phe obtained by soaking (3KGF) showed that upon simultaneous binding of Trp and Phe, there are two Phe molecules bound in the tight dimer interface and two Trp molecules bound in the tetramer interface. A secondary Phe site can also be observed, however structure of Trp+Phe co-crystallised with *Mtu*DAH7PS (3RZI) suggested that this secondary Phe binding site is not important for inhibition of *Mtu*DAH7PS by the Trp+Phe combination. Tyr molecules are shown to bind in the same binding sites as Phe molecules (as observed in Tyr+Trp and Tyr-only structures). One extra hydrogen bond is formed between Tyr molecule and residues in the secondary Phe binding site, indicating that the "secondary Phe binding site" may have higher affinity for Tyr molecules.

Crystal structure of *Mtu*DAH7PS in complex with single aromatic amino acids showed that the allosteric ligand binding sites can be quite flexible, at high ligand concentrations, Phe molecules are observed to bind in the Trp binding sites at the tetramer interface, and Trp molecules are observed to bind in both the primary Phe binding sites in the dimer interface and the secondary Phe binding sites.

Comparison of the structures of *Mtu*DAH7PS in complex with different aromatic amino acids shows clearly that there is no gross conformational change within *Mtu*DAH7PS associated with any ligand binding event. This suggests that mechanism of allosteric regulation and the communication among distant binding sites of *Mtu*DAH7PS is mainly contributed by alteration of protein dynamics upon ligand binding. Therefore, MD simulations can be employed to study the molecular basis of this usual synergistic allosteric regulation of *Mtu*DAH7PS (discussed in Chapter 6).

Chapter 5. Investigating the active site: molecular modelling of active site inhibitors into *Mtu*DAH7PS

5.1 Introduction to mechanism of reaction catalysed by DAH7PS

DAH7PS catalyses the first reaction in the shikimate pathway, which is an aldol condensation between PEP and E4P to form the seven carbon sugar DAH7P (Figure 5. 1).

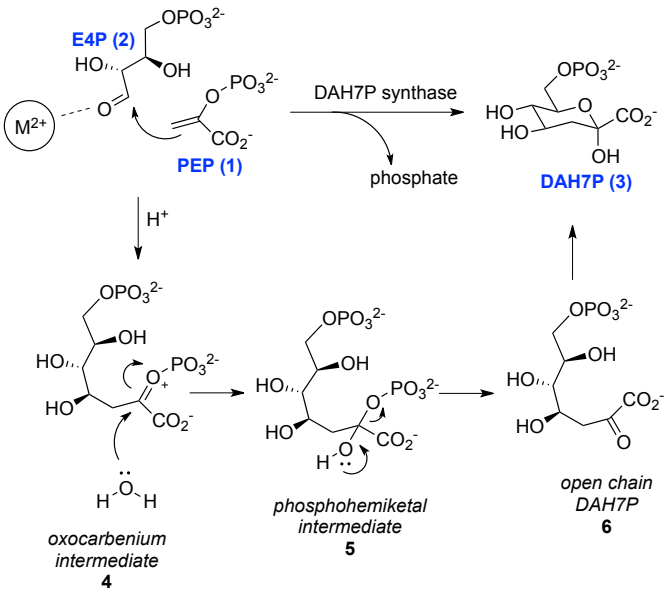


Figure 5. 1. Overall reaction catalysed by DAH7PS and proposed mechanism of condensation of PEP with E4P.

Many of the details of the reaction mechanism catalysed by DAH7PS are known, and crystal structures of the enzymes from several sources have been solved, revealing several features of the reaction intermediates that could be targeted in inhibitor design (*57, 59, 61-63, 67*). The reaction occurs stereospecifically with respect to both substrates, with the *si* face of PEP attacking the *re* face of E4P (*79-81*). A divalent metal ion present in the active site is essential for activity. The initial attack of PEP on E4P is believed to be promoted *via* Lewis acid activation of the E4P aldehyde group by a divalent metal ion, followed by attack of PEP to give an oxocarbenium

intermediate **4** with trigonal planar geometry at the former C2 of PEP (Figure 5. 1) (*61, 82*). Attack of water on the electrophilic oxocarbenium ion gives rise to tetrahedral phosphohemiketal intermediate **5**, and elimination of phosphate leads to the open chain form **6** of the product DAH7P. For the step of water attack, the water molecule can potentially attack from either face of **4**, giving rise to two possible diastereoisomers of tetrahedral intermediate **5**, differing in their absolute configuration at C2. Although this stereogenic centre is transient, and the stereochemical information is lost by elimination of phosphate in the final step to generate the product DAH7P, the geometry of the enzyme active site is likely to favour stereoselective attack of water to form one diastereoisomer of **5**. The DAH7P is formed in its acyclic form, and cyclises into the cyclic pyranose form following release from the enzyme.

In this chapter, the substrates (PEP and E4P) and the intermediate molecule (**5**) were modelled into the active site of *Mtu*DAH7PS, to understand the roles the active site residues of DAH7PS play in substrates binding and catalysis, and to help identify the preferred side of water attack on the oxocarbenium species (**4**) during the reaction. This chapter also describes the modelling studies conducted on a series of intermediate mimicking inhibitors, in order to support existing results from synthetic work, and to guide future design of inhibitor molecules for *Mtu*DAH7PS.

5.2 Docking studies for *Mtu*DAH7PS

5.2.1 Docking E4P into Mtu*DAH7PS active site*

To understand the binding mode of the two reaction substrates, and to identify the active site residues involved, the substrate E4P was docked into the active site of *Mtu*DAH7PS. The crystal structure of the ligand-free *Mtu*DAH7PS (PDB ID 2B7O) was used as the starting coordinates to develop the model of the active site receptor for the docking of E4P. One of the two substrates, PEP, was already bound in the active site of this crystal

structure. A grid file of the receptor was generated for each docking calculation to increase computation efficiency. The shape and properties of the receptor are represented on the grid file. In order to generate the grid file for docking E4P, the structure of the active site was first prepared by adding the missing side chains of residues that were not modelled in the original crystal structure due to disorder. Hydrogen atoms were also added because they were not present in the crystal structures. The structure was then minimised in MacroModel using OPLS_2005 force field and the PRCG method, with water as the solvent (*38*), and residues within 5 Å of the active site Mn^{2+} ion were "frozen" during the minimisation. The receptor grid was generated based on the minimised structure using Glide (*83*), with the grid centre defined as the centroid of residues 284, 133, 134, 135, 440, 441 and the active site PEP, the size of the grid was defined with the option "docking ligands shorter than 20 Å". The substrate E4P was then docked in to this rigid receptor model using Glide (*83*).

5.2.1.1 Docking E4P into MtuDAH7PS active site in which one water molecule is bound to Mn^{2+}

In the original crystal structure of *Mtu*DAH7PS (PDB 2B7O), one water molecule (HOH2024) is coordinated to the active site metal ion. In the first docking attempt, this water molecule was retained for the generation of the grid file, and the position of the water molecule was kept rigid during the docking calculation The docking produced six poses of E4P in the active site. The pose with the best docking score is shown in Figure 5. 2. The E4P molecule is held in the active site by forming hydrogen bonds with Pro134 backbone carbonyl and the side chain of Lys133. The aldehyde of E4P forms hydrogen bond with Arg248 which is bridging between the two substrates PEP and E4P. As a result of this hydrogen bonding, the aldehyde exposes its *si* face to the attacking PEP carbon in this pose. None of the poses obtained contain E4P aldehyde with an orientation consistent to the known experimental results which require *re* face attack of PEP (all poses have *si* face facing PEP C3). Furthermore, the distances between the two reacting atoms in all six poses are greater than 3.5 Å, which places the two substrates

too far away for reaction to occur.

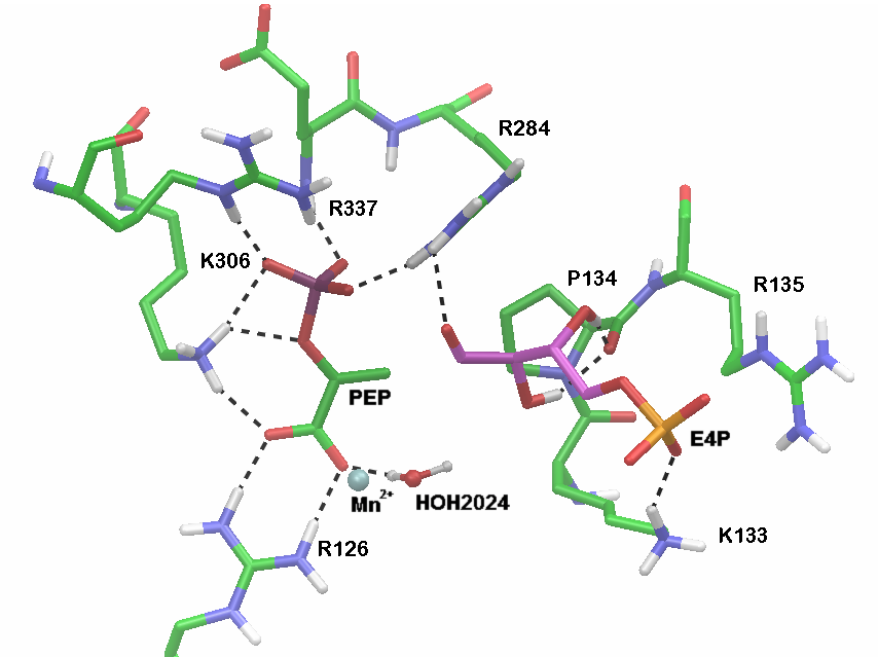


Figure 5. 2. Best scored pose of E4P in *Mtu*DAH7PS active site with one water molecule bound to Mn^{2+}. The active site residues and the bound PEP are shown with green carbon atoms and the docked E4P is shown with magenta carbon atoms. The active site metal ion (Mn^{2+}) is shown as a cyan sphere. Hydrogen bonds are displayed as black dashed lines.

5.2.1.2 Docking E4P into MtuDAH7PS active site with no water molecule bound to Mn^{2+}

The reason why docking results are inconsistent with experimental results could be due to the water molecule (HOH2024) that was left in the active site during docking. This water molecule occupies the space near the active site metal ion, and as a result, the E4P molecule cannot get closer to either the PEP molecule or the Mn^{2+} ion. Therefore, the water molecule was removed from the active site, and the receptor grid regenerated. The E4P molecule was then docked into the new grid file.

The best pose from this second attempt of docking was selected to be the best scored pose that has the *re* face of E4P aldehyde facing the PEP molecule. In this best pose (Figure 5. 3), PEP forms a network of hydrogen bonds with the active site residues, including Arg126, Lys306, Arg337, Glu283, and Arg284. These hydrogen bonds formed are retained from the original crystal structure

(2B7O). The two hydroxyl groups on the E4P molecule form hydrogen bonds with the backbone carbonyl of Pro134, and the phosphate group on E4P forms hydrogen bonds with Lys133 and Lys380. The E4P molecule is held in such a way that the *re* face of the aldehyde is facing C3 of PEP, which is consistent with the experimental finding that attack occurs on the *re* face of the aldehyde. However, the distance between C3 of PEP and the aldehyde carbon on E4P in this model is still quite long at 3.5 Å.

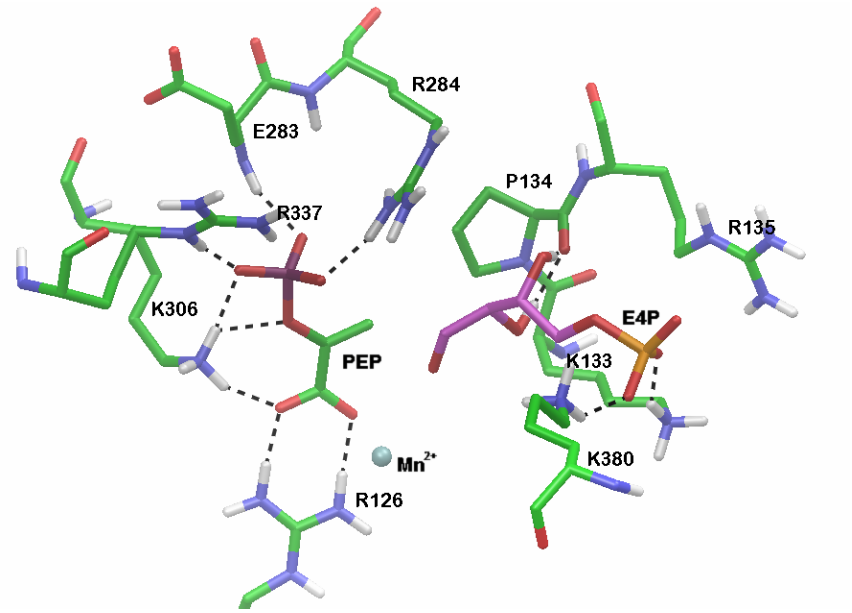


Figure 5. 3. Best pose from rigid docking of E4P in the active site of ligand-free *Mtu*DAH7PS (PDB 2B7O) with no water bound to Mn^{2+} ion. Hydrogen bonds are shown as black dashed lines.

The position of the phosphate group in E4P in the model is different from that of the sulfate ion found in the active site of the original crystal structure of the ligand-free *Mtu*DAH7PS (PDB 2B7O). A superposition between the original crystal structure and the modelled structure reveals this difference (Figure 5. 4). The distance between the sulfur atom in the sulfate ion and the phosphorus atom in the phosphate group on E4P is 4.8 Å. It was proposed that the position of the sulfate ion in the crystal structure indicates the likely position of the phosphate group of E4P in the active site (*67*). In the crystal structure, the sulfate ion forms a salt bridge with the nearby Arg135 residue, and a hydrogen bond with side chain of Ser136. However, the best pose of E4P predicts the phosphate group to be positioned between positively charged side chains of residues Lys133 and Arg135 and Lys380. The

negative charge on the phosphate is stabilised in the positively charged environment.

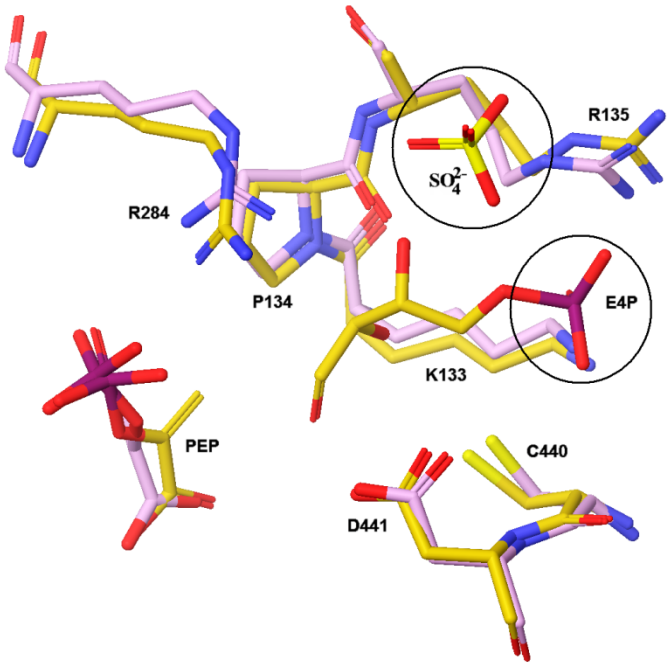


Figure 5. 4. Superposition of the active site of ligand-free *Mtu*DAH7PS from the crystal structure 2B7O and the modelled structure from rigid docking. The crystal structure is displayed with pink carbon atoms, and the modelled structure with yellow carbon atoms. The two circles highlight the sulfate ion in the crystal structure and the phosphate group of E4P in the modelled structure.

5.2.2 The induced fit docking protocol

In standard virtual docking studies, ligands are docked into the binding site of a receptor where the receptor is held rigid and the ligand is free to move. However, the assumption of a rigid receptor can give misleading results, since in reality many proteins undergo side chain or backbone movements, or both, upon ligand binding. These changes allow the receptor to alter its binding site so that it more closely conforms to the shape and binding mode of the ligand. To allow for a change in receptor conformation upon ligand binding an "induced fit" protocol has been incorporated into the modelling software (*84*).

The induced fit docking protocol (*84*) surmounts the shortcoming of rigid docking, by allowing movement of active site residues close to the docked

ligand, thereby accounting for the "induced fit" effect on the receptor upon ligand binding. In this procedure, the ligand molecule is initially docked into the receptor with scaled atomic van der Waals radii, then the side chain conformations of the active site residues close to the docked ligand (within 5 Å) are optimised to form the best interactions with the ligand molecule. The optimised side chain conformations of residues represent the "induced fit" of the active site to the docked ligand. In the final step, the ligand molecule is re-docked into the "induced-fit" conformation of the active site and the docked pose is evaluated and examined. The induced fit docking therefore can produce more accurate prediction of ligand binding mode, especially for large ligand molecules that usually do not model well with rigid docking.

5.2.3 Induced fit docking of reaction intermediate

The problem resulting from rigid docking of E4P into the active site receptor with PEP already bound is that in all the docked poses the distance between the two reacting atoms are too great for reaction to occur. This is because the van der Waals interaction becomes unfavourable when any two atoms get too close to each other. However, for the reaction to occur, C3 of PEP and C1 of E4P have to approach each other, as a result, the rigid docking procedure is not predicting the position of E4P before reaction happens.

To overcome the distance problem, the predicted tetrahedral intermediate **5** (Figure 5. 1) of the reaction between PEP and E4P was docked into the active site of the ligand-free *Mtu*DAH7PS. The position of the aldehyde of E4P and C3 of PEP could then be predicted based on how the intermediate is held in the active site. Since it is unclear which side of the oxocarbenium species (**4**) the water attacks from, both possible diastereoisomers of this tetrahedral intermediate (**5**) were docked into the active site receptor.

The structures of both possible diastereoisomers of the tetrahedral intermediate were built using the protein builder in Maestro (*39*) of the Schrödinger Suite 2006 and then minimised with MacroModel (*38*). The Polak-Ribiere Conjugate Gradient (PRCG) minimisation method was used

with up to 5000 iterations and a gradient convergence threshold of δ=0.05 kJ/(mol*Å). A conformational search was then carried out for each isomer in order to generate an ensemble of low-energy conformers to find a suitable starting conformation for the later docking calculation. Conformational searches were conducted with the MCMM Serial Torsional Sampling method and were run with a GB/SA water model, using the OPLS2005 force field, with 3000 steps for the conformational search and an energy window of 12.0 kJ/mol for collecting appropriate conformers. All the conformers from the conformational search were clustered by consideration of torsional RMS of the backbones, and the representative conformations from the clustering analysis were selected based on the clustering statistics, which were then used in the docking calculations as the starting conformations for each diasteroisomer of the tetrahedral intermediate.

The modelling of the tetrahedral intermediate into the active site of *Mtu*DAH7PS was carried out with the Schrödinger Suite 2006 Induced Fit Docking protocol (*84*). The active site in chain A of the ligand-free crystal structure of *Mtu*DAH7PS (2B7O) was used as the receptor. The centre of the receptor grid was defined as the centroid of residues 87, 126, 133, 134, 135, 136, 137, 248, 250, 280, 282, 283, 284, 306, 337, 366, 369, 380, 382, 409, 411, 440, and 441. For the initial docking, the van der Waals radii of the linear intermediate atoms were scaled to 0.8. The 20 best solutions of initial docking were kept. All residues of the protein within a 5 Å distance of the respective linear molecule solution were refined. The ligands were re-docked, with a van der Waals radii scale of 0.8, to the top 20 newly generated protein structures if the energy was within 30 kcal/mol of the best protein structure.

This docking study produced a total of 30 poses for both diastereoisomers and out of which 17 poses showed retained PEP binding interactions. All of the 17 poses with the retained PEP orientation are from the *S*-isomer of the tetrahedral intermediate. This suggests that the *R*-isomer does not fit in the active site of *Mtu*DAH7PS as well as the *S*-isomer. The *S*-isomer corresponds to the tetrahedral intermediate molecule if the water attacks obxocarbenium species from the opposite side of the active site Mn^{2+} ion (i.e.

the *re* face), therefore, the modelling results suggest the direction of the incoming water molecule is from the *re* face of the oxocarbenium intermediate (**4**).

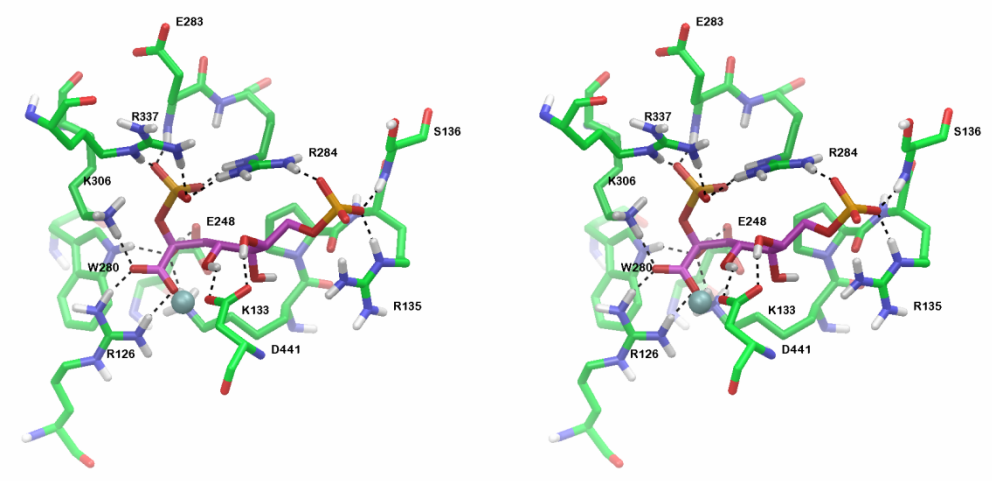


Figure 5. 5. Stereo view of the best scored pose of linear intermediate (*S*-isomer) from induced fit docking. The active site carbons are displayed in green, and the docked intermediate is shown with magenta carbon atoms, and the active site metal ion is displayed as the cyan sphere. Hydrogen bonds are represented by black dashed lines.

The best pose from this induced fit docking (Figure 5. 5) shows a network of hydrogen bonds between the tetrahedral intermediate and the enzyme active site. The PEP part of the intermediate forms hydrogen bonds between the carboxylic acid group and Arg126, this interaction is also observed in the ligand-free crystal structure of *Mtu*DAH7PS (PDB 2B7O). Lys306 also interacts with the negative carboxylic acid on PEP. The phosphate group on PEP is forming hydrogen bonds and salt bridges with the nearby Arg337 and Arg284 and one hydrogen bond with the backbone NH group of Glu283. In this pose, bridging between the PEP part and the E4P part of the intermediate is observed. This bridging occurs via Arg284 by forming salt bridges and hydrogen bonds with the two phosphate groups on PEP and E4P. The E4P phosphate group is held by side chains of both Arg284 and Arg135, and the backbone NH group of Ser136. This indicates that Arg135 can play a role in positioning of the E4P phosphate group. The previous E4P aldehyde, now the newly formed OH group on the linear intermediate, is at a close distance (2.2 Å) to the active site Mn^{2+} ion, and this suggests metal coordination to the Mn^{2+} ion via the oxygen atom may occur. This indicates that the aldehyde of

E4P may be held in place by coordination to the metal ion before reaction occurs. The OH group that corresponds to the incoming water molecule is forming hydrogen bonds with Trp280, Glu248, and Lys133. This suggests that the incoming water molecule may be held in place in the active site before reaction occurs by the above residues. Two of these residues, Glu248 and Lys133, are conserved across the entire DAH7PS family, whereas Trp280 is only conserved in the type II enzymes.

One of the problems of the results from rigid docking of E4P into the active site (section 5.2.1) is that the position of the E4P phosphate was never predicted to be close to the sulfate ion that was found in the active site of the original crystal structure (2B7O). In order to check whether the induced fit docking of the linear intermediate can predict such a position for the phosphate moiety, the best pose from induced fit docking was superimposed with the ligand-free crystal structure (2B7O) with the sulfate ion bound (Figure 5. 6). The distance between the predicted position of E4P phosphate and the bound sulfate ion is 1.8 Å, which is a significant improvement, compared to the best pose from rigid docking of E4P (4.8 Å, Figure 5. 4).

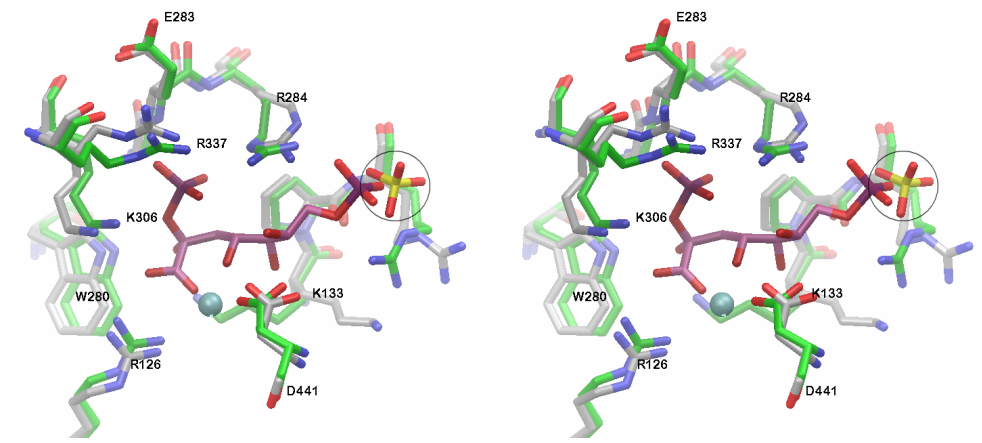


Figure 5. 6. Stereo view of the superposition between the best pose from the induced fit docking of the linear intermediate molecule (green carbon atoms for residues at the active site) and the original crystal structure (2B7O, grey carbon atoms). The sulfate ion present in the original crystal structure is highlighted in black circle, and the docked linear intermediate is shown with purple carbon.

The best pose was then superimposed on the crystal structure of *Sce*DAH7PS (Figure 5. 7), and showed that the position of the E4P phosphate group in the

modelling pose is also quite close to the position of the G3P phosphate group in the crystal structure, which was considered to be the best indication of the position of E4P phosphate. The distance between the predicted E4P phosphate and the G3P phosphate is 1.6 Å.

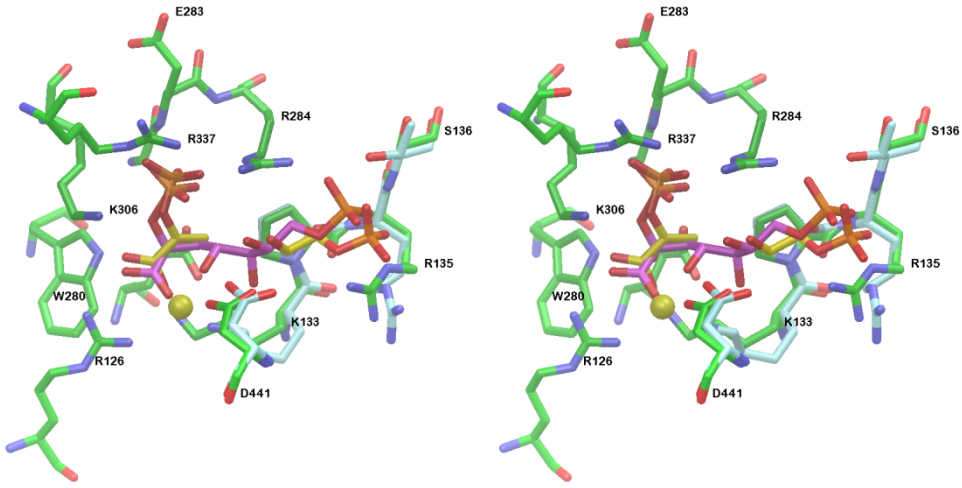


Figure 5. 7. Stereo view of the superposition of the best pose of the linear intermediate from induced fit docking (green carbons) with the crystal structure of *Sce*DAH7PS (PDB ID 1OF8) (cyan carbons) with G3P bound in the active site. PEP and G3P in *Sce*DAH7PS are shown with yellow carbon atoms.

One of the interesting features in the modelling result is the position of the Lys133 side chain. The Lys133 residue in the original *Mtu*DAH7PS crystal structure (2B7O) is found in a completely different conformation (Figure 5. 8B) from the docking pose. A different Lys133 conformation is also observed in the crystal structure of *Sce*DAH7PS (Figure 5. 8A). This large difference in side chain conformations for the absolutely conserved Lys133 shows the extent of flexibility of this residue. In the modelled pose for the tetrahedral intermediate, the Lys133 residue forms a hydrogen bond with the oxygen atom on the incoming water molecule. This suggests that the Lys133 side chain may help position the incoming water molecular and play a role in facilitating proton transfer during the reaction.

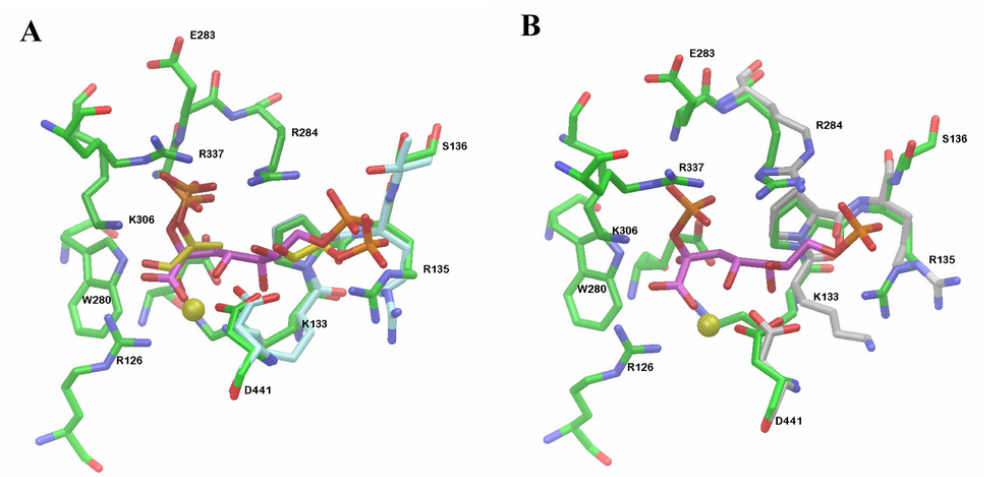


Figure 5. 8. Superposition between active site residues of *Mtu*DAH7PS in the best pose from induced fit docking (displayed in green carbons) and (A) the crystal structure of *Sce*DAH7PS (PDB ID 1OF8, displayed in cyan carbons); (B) the original crystal structure (PDB ID 2B7O, displayed in grey carbons). The active site metal ions are displayed as yellow spheres.

5.2.4 Docking of intermediate mimicking inhibitors

Based on the results from induced fit modelling of the tetrahedral intermediate molecule **5**, a series of inhibitors mimicking the structure of the linear intermediate was designed by Sebastian Reichau from our research group, and here molecular modelling technique was used to predict the binding mode and potency of these structures as potential inhibitors. In order to carefully evaluate results from docking calculations, two different receptor grids were generated.

The first grid was based on the wild type crystal structure of *Mtu*DAH7PS (PDB ID 2B7O). The grid file was generated with Glide. The centre of the grid was defined as the centroid of residues 87, 126, 133, 134, 135, 136, 137, 248, 250, 280, 282, 283, 284, 306, 337, 366, 369, 380, 382, 409, 411, 440, and 441.

A second grid was generated based on the best pose from induced fit modelling of the tetrahedral intermediate above. The active site conformation, especially the residues in close proximity to the modelled intermediate molecule (within 5 Å), adopted different conformations from the original crystal structure (2B7O), to optimise interactions with the modelled

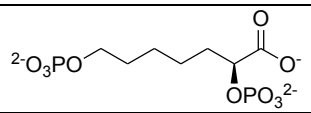
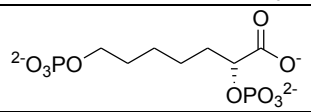
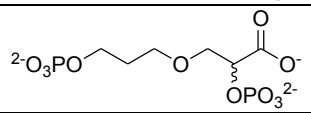
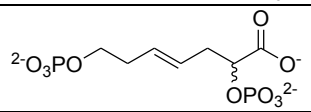
tetrahedral intermediate. The active site conformation can be considered as "intermediate-adapted". This particular active site conformation may be useful in modelling of the following compounds, since these molecules were designed to mimic the intermediate molecule. The centre of the grid was defined as the centroid of the modelled intermediate molecule.

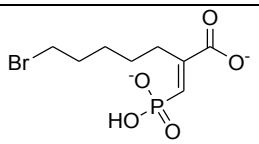
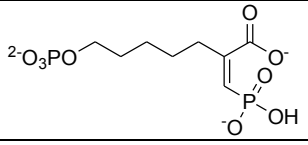
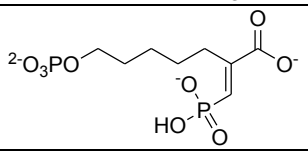
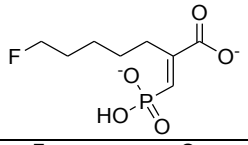
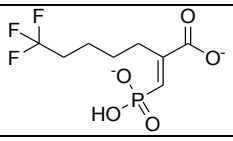
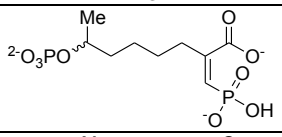
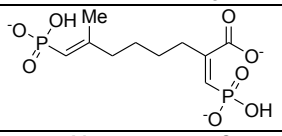
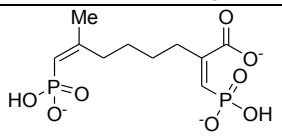
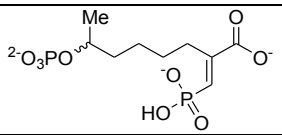
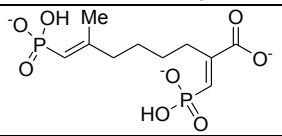
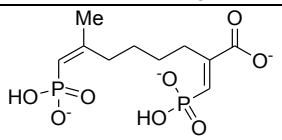
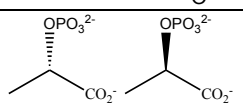
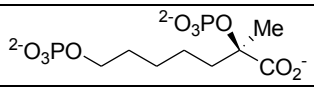
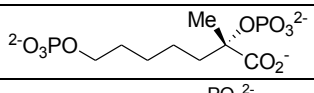
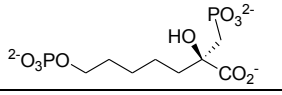
The series of intermediate mimicking structures (

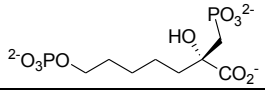
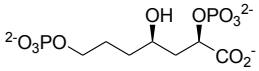
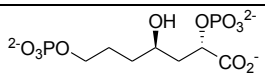
Table 5. **1**) were treated the same way as that in the modelling of the tetrahedral intermediate. The structures were first minimised in MacroModel, and then a conformational search was conducted to search for low energy conformers. The low energy conformations of each structure were then clustered and the selected representative conformations for each structure obtained from this clustering analysis were modelled into both receptor grids with rigid docking procedure and XP (extra precision) mode with Glide. In general each structure had a better docking score in the receptor with intermediate adapted conformation (

Table 5. **1**), suggesting that this receptor is better suited for this series of designed structures. Therefore, docking poses from the intermediate adapted receptor were mainly considered for evaluation below.

Table 5. 1. Table of structures and docking scores for the series of designed intermediate-mimicking structures modelled in two different receptor grids.

Number	Structure	Docking scores in the wild type receptor grid (based on crystal structure 2B7O)	Docking scores in the intermediate adapted receptor grid
S-5.1		-5.11	-5.81
R-5.1		-4.84	-5.96
S-5.3 *R*-5.3		-4.95 -4.69	-5.91 -5.94
S-5.4 *R*-5.4		-5.07 -5.01	-6.60 -8.26

5.5		-4.74	-5.76
5.6		-5.29	-6.37
5.7		-5.33	-6.53
5.11		-4.44	-5.51
5.12		-4.33	-5.44
S-5.13 *R*-5.13		-5.27 -5.75	-5.11 -6.27
5.14		-5.51	-5.66
5.15		-5.64	-6.04
S-5.16 *R*-5.16		-5.69 -5.79	-6.54 -5.77
5.17		-6.18	-6.36
5.18		-7.21	-6.68
S-5.19 *R*-5.19		-3.55 -3.63	-4.98 -5.37
R-5.30		-5.35	-7.62
S-5.30		-5.14	-7.58
R-5.31		-5.12	-7.77

S-5.31		-6.77	-6.50
R-5.32		-4.82	-6.63
S-5.32		-4.97	-7.33

Compound **5.1** (Figure 5. 9) is the simplified analogue of the intermediate molecule with the hydroxyl groups removed. Both stereoisomers of **5.1** were modelled and the results showed that the *R*-isomer of **5.1** has slightly better docking scores than the *S*-isomer. However, both isomers adopt similar binding modes, which are also very similar to those of the linear intermediate in the active site (Figure 5. 5). The phosphate moieties in the PEP mimicking portion of both isomers of compound **5.1** form interactions with Arg337, Lys306, Arg126 and Arg284. The phosphate moieties of the E4P mimicking portion of both isomers form interactions with Arg135 and the backbone of Ser136. Unlike the selectivity observed in the docking of the actual tetrahedral intermediate, where only the *S*-isomer can fit in the active site, here both the *S*- and *R*- isomers of **5.1** can fit in the active site with similar binding mode. This is probably due to the absence of a hydroxyl group at the C2 centre.

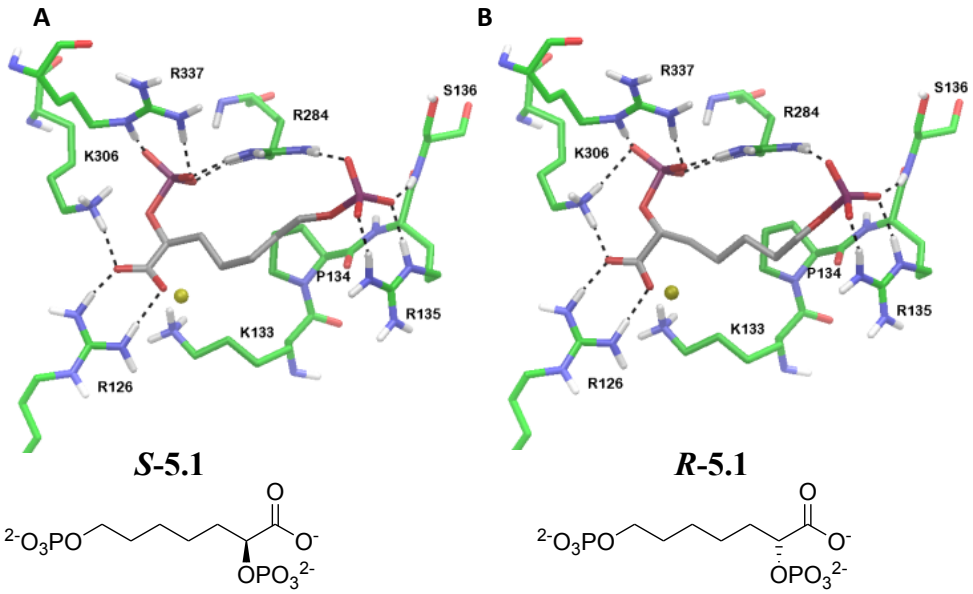


Figure 5. 9. Best pose from docking compounds *S*-isomer (A) and *R*-isomer (B) of compound 5.1 into intermediate adapted receptor of *Mtu*DAH7PS.

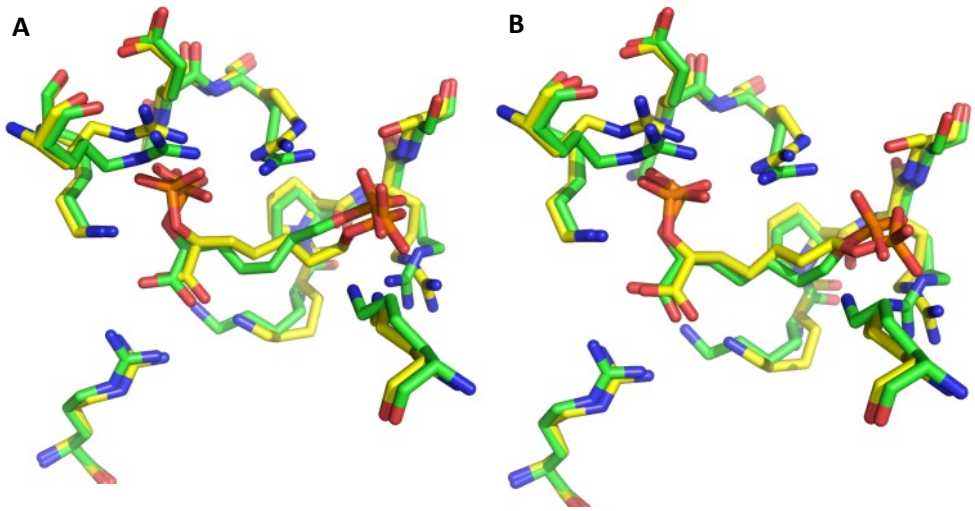


Figure 5. 10. Superposition between structure of *Mtu*DAH7PS co-crystallised with racemic mixture of compound 5.1 (PDB 3PFP) and the modelling poses for compound 5.1s (A) and 5.1r (B). The crystal structure is displayed with yellow carbon atoms and the modelling poses are displayed with green carbon atoms.

A structure of *Mtu*DAH7PS co-crystallised with the racemic mixture of compound **5.1** was obtained and refined at 2.35 Å (PDB 3PFP) by Sebastian Reichau (*85*). In this crystal structure, based on the electron density maps obtained, it could not be concluded whether there is preferential binding of one of the diastereroisomers of **5.1** from the racemic mixture, therefore both *R*- and *S*-isomers of **5.1** were included in the structure. Inhibition studies (also carried out by Sebastian Reichau) conducted with pure *S*- and *R*-isomers of compound **5.1** showed that both isomers have a high affinity to the enzyme active site, with the *R*-isomer have slightly higher affinity, consistent with the *R*-isomer having slightly better docking score in the modelling study. The interactions formed between **5.1** and the enzyme active site in the crystal structure are the same as those predicted in the modelling results for **5.1** (Figure 5. 10), and the inhibitors adopt binding modes nearly identical to the modelled pose. Interestingly the movement of Lys133 side chain that was predicted in the induced fit docking of the linear intermediate is also observed in this crystal structure. This validates that the modelling approach used here is effective for predictions of enzyme-inhibitor interactions for *Mtu*DAH7PS, and the intermediate adapted receptor conformation is more suitable for predicting the potency of the series of intermediate-mimicking inhibitors.

5.2.5.1 Docking of intermediate mimicking structures that have not been synthesised

The structures discussed below have not yet been synthesised at the time of writing this thesis. All modelling results and the predictions of the potencies of the designed molecules should be considered as a guide to improving inhibitor design, and for filtering out the unnecessary synthesis of compounds that likely will prove to be weak inhibitors (tables showing number of contacts made in each of the modelled pose are in Appendix 2 and 3, and a plot of different components of the docking scores for each modelled structure is in Appendix 4).

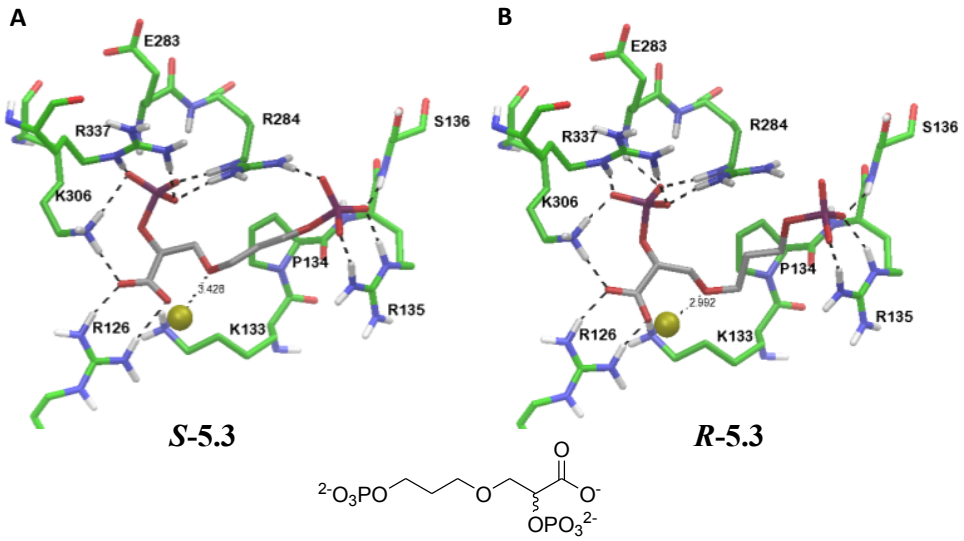


Figure 5. 11. Best poses from docking the *S*-isomer (A) and *R*-isomer (B) of structure 5.3 into the intermediate-adapted receptor conformation of *Mtu*DAH7PS.

Structure **5.3** was designed with an oxygen atom inserted in the chain to coordinate to the active site metal ion (Mn^{2+}) and hence to increase binding affinity. However, docking both isomers of this compound into the intermediate adapted receptor suggests that the binding affinity may not be greatly improved by the presence of the inserted oxygen (Figure 5. 11). The docking scores of **5.3** are very close to the simple intermediate mimic **5.1**. The insertion of oxygen atom seems to reduce the number of hydrophobic contacts with the binding pocket of the active site and no extra hydrogen bonding interactions are formed. The distance between the inserted oxygen

atom and the active site metal ion is 3.4 Å and 3.0 Å for the *S*-isomer and *R*-isomer of **5.3** respectively, which is slightly long for coordination to the metal ion.

Although modelling did not predict metal coordination between the inserted oxygen atom in compound **5.3** and the active site metal ion, it may still occur in reality, due to the limitation of the docking method. In the docking procedure used here, the metal ion in the active site was treated as a point charge with a defined atomic radius and interactions with the electron orbitals of the metal ion that are essential for predicting coordination were not considered during the docking calculation. Therefore, metal coordination cannot be predicted with any accuracy in the current docking method. Whether this inhibitor is potent, and whether it forms coordination to active site metal ion as intended should be verified by experiments.

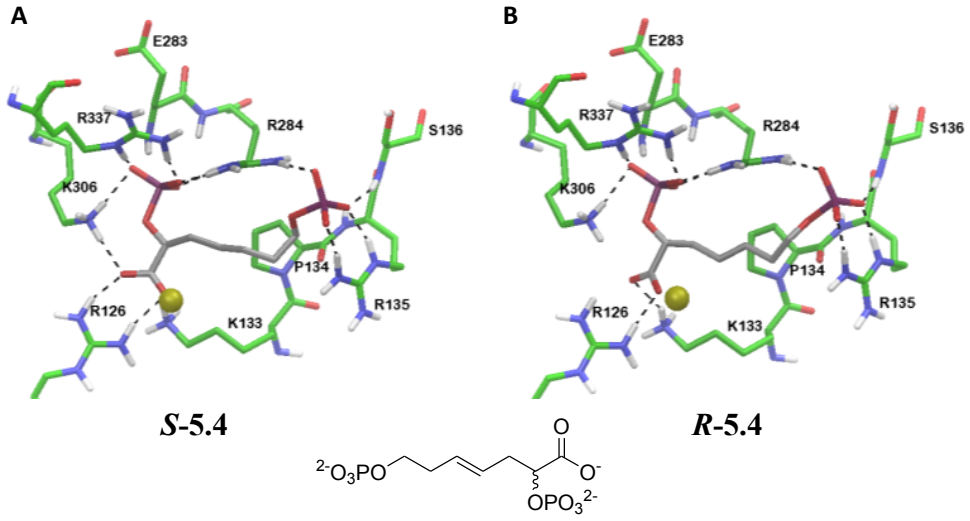
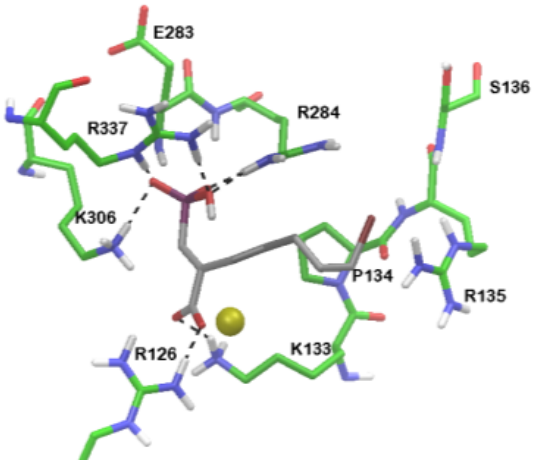
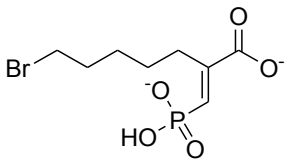


Figure 5. 12. Best poses from docking the *S*-isomer (A) and *R*-isomer (B) of structure 5.4 into the intermediate-adapted receptor conformation of *Mtu*DAH7PS.

Structure **5.4** contains a double bond between C4 and C5 of the carbon chain (Figure 5. 12). Both isomers of this structure are less bulky around the double bond in comparison to compound **5.1**, which could give both isomers fewer close contacts and better van der Waals score. The carboxylate group in the PEP-mimicking portion of the *R*-isomer is twisted slightly so instead of forming two hydrogen bonds with the side chain of Arg126 as is observed for

the other inhibitors above, one hydrogen bond is formed with the side chain of Lys133.





5.5

Figure 5. 13. Best poses from docking the structure 5.5 into the intermediate-adapted receptor conformation of *Mtu*DAH7PS.

Structure **5.5** (Figure 5. 13) is slightly different from the above inhibitors, as it contains a Br atom, which replaces the phosphate group at the E4P end of the molecule. As expected this compound does not form hydrogen bonds and salt bridges between the E4P end of the inhibitor and the enzyme active site. However, this molecule is predicted to fit well with the shape of the binding pocket, with very good hydrophobic contacts to the active site. This suggests that despite the lack of hydrogen bonding interactions with the enzyme active site, structure **5.5** may still be a potent inhibitor for *Mtu*DAH7PS.

Structures **5.6** and **5.7** are geometric isomers at the PEP-mimicking portion (Figure 5. 14). Structure **5.6** (the *cis*-isomer) has better van der Waals scores than **5.7** (the *trans*-isomer). However, this is only due to the model predicting that the PEP-mimicking portion of **5.6** is not binding as deep in the pocket as in structure **5.7**. Structure **5.7** can also pick up more hydrogen bonds at the PEP-mimicking portion of the structure, indicating it may be more potent than **5.6**.

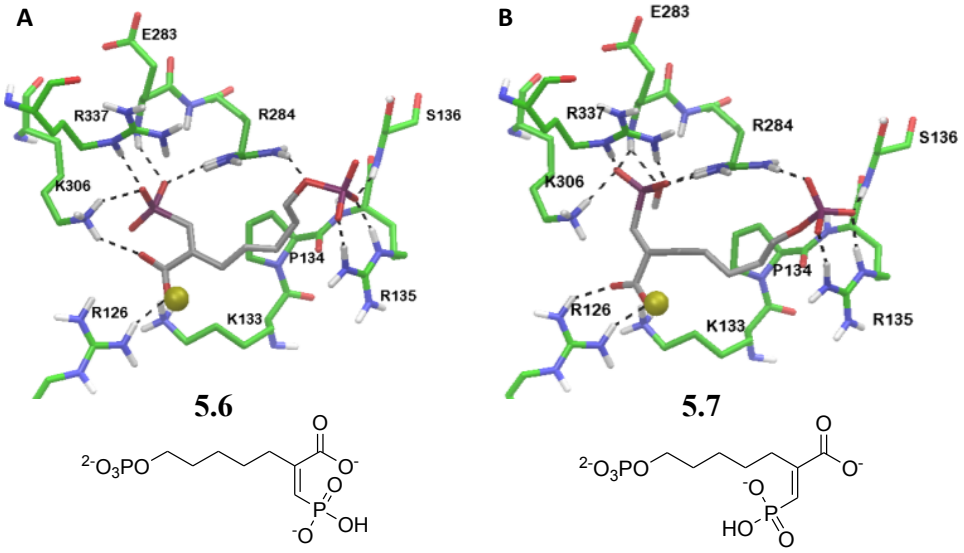


Figure 5. 14. Best poses from docking structure 5.6 (A) and 5.7 (B) into the intermediate-adapted receptor conformation of *Mtu*DAH7PS.

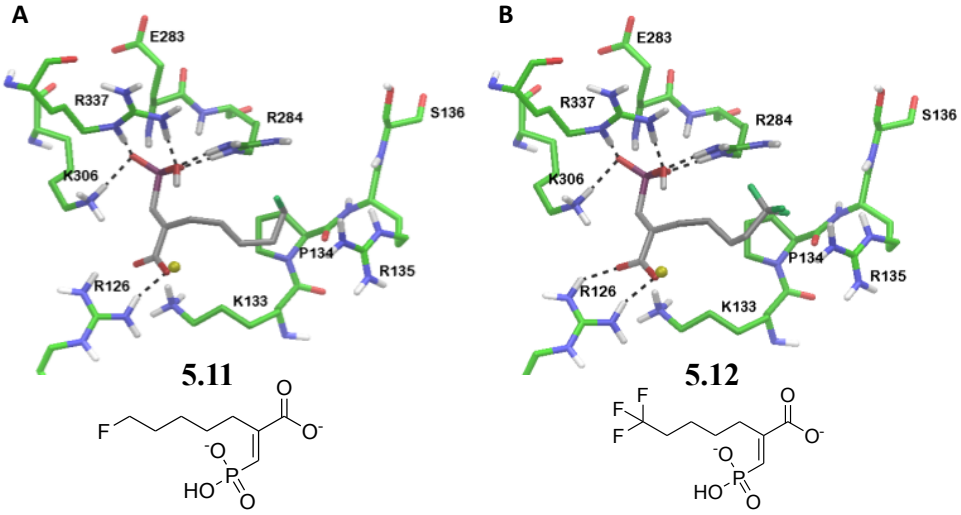


Figure 5. 15. Best poses from docking structure 5.11 (A) and 5.12 (B) into the intermediate-adapted receptor conformation of *Mtu*DAH7PS.

Structures **5.11** and **5.12** are analogues of **5.5** (with the Br atom at the E4P end) (Figure 5. 15). The PEP-mimicking portions of both **5.11** and **5.12** adopt the same binding mode as in **5.5**. However, the favourable hydrophobic score that is in **5.5** is disrupted by replacing Br with the more electronegative F. The van der Waals interactions in **5.11** and **5.12** in the model show no improvement compare to **5.5**.

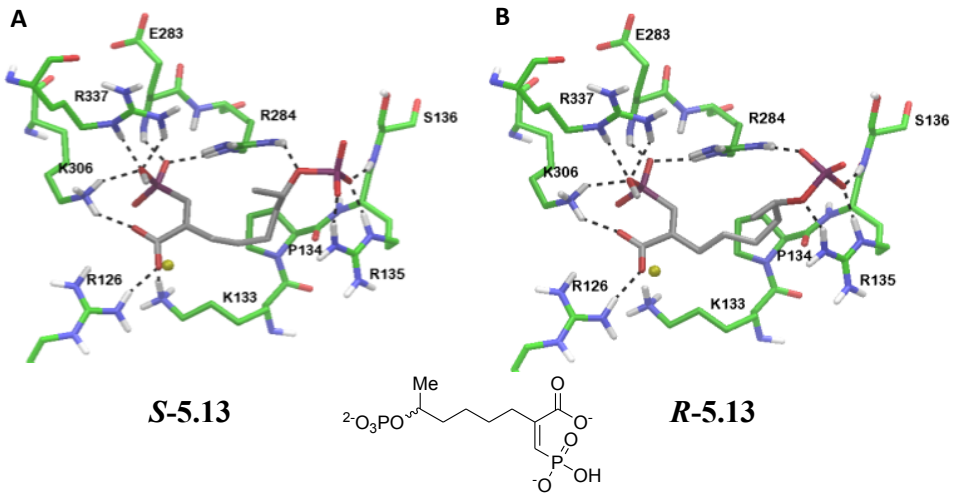


Figure 5. 16. Best poses from docking the *S*-isomer (A) and *R*-isomer (B) of structure 5.13 into the intermediate-adapted receptor conformation of *Mtu*DAH7PS.

Structure **5.13** has *cis* geometry in the PEP-mimicking portion. Modelling shows that both isomers of this structure can bind in the active site of *Mtu*DAH7PS (Figure 5. 16). Results suggest that this structure has fewer close contacts with the active site residues but this is only due to the PEP-mimicking portions of the isomers not binding as deep in the active site pocket as the intermediate molecule (or compound **5.1**). The E4P-mimicking parts of **5.13** isomers possess an added methyl group to probe any hydrophobic space in the active site. However, modelling suggests that the added methyl group will not add many extra hydrophobic contacts.

Structures **5.14** and **5.15** both have the same PEP-mimicking portion as **5.13**, with the *cis* geometry, and as expected the PEP-portions of these two compounds bind in the same way as in **5.13** (Figure 5. 17). The E4P-mimicking portions of **5.14** and **5.15** are modified based on **5.13** with a double bond introduced to restrict the orientation of the added methyl group. Modelling results suggest that the double bond will not increase potency of the inhibitor, especially in compound **5.14**, where the methyl group is forced to point outside the active site pocket, to the solvent exposed space.

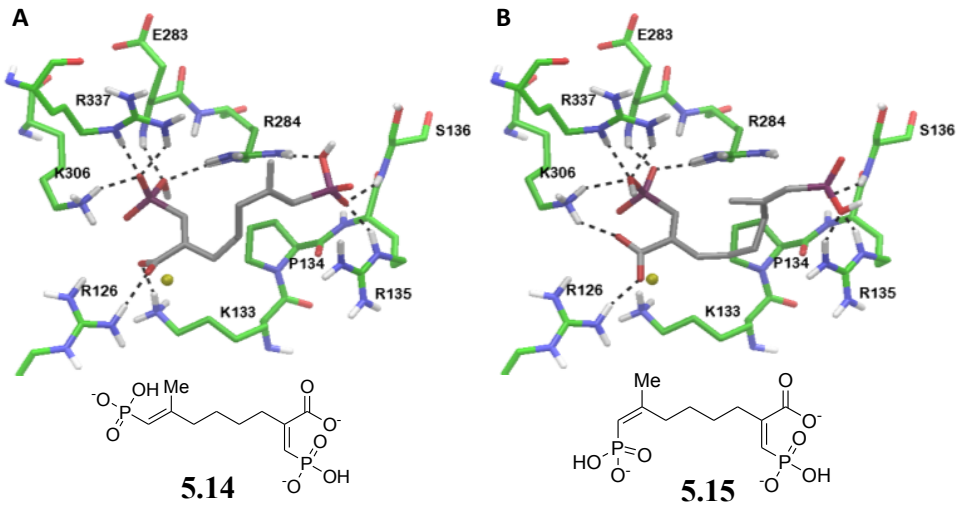


Figure 5. 17. Best poses from docking structure 5.14 (A) and 5.15 (B) into the intermediate-adapted receptor conformation of *Mtu*DAH7PS.

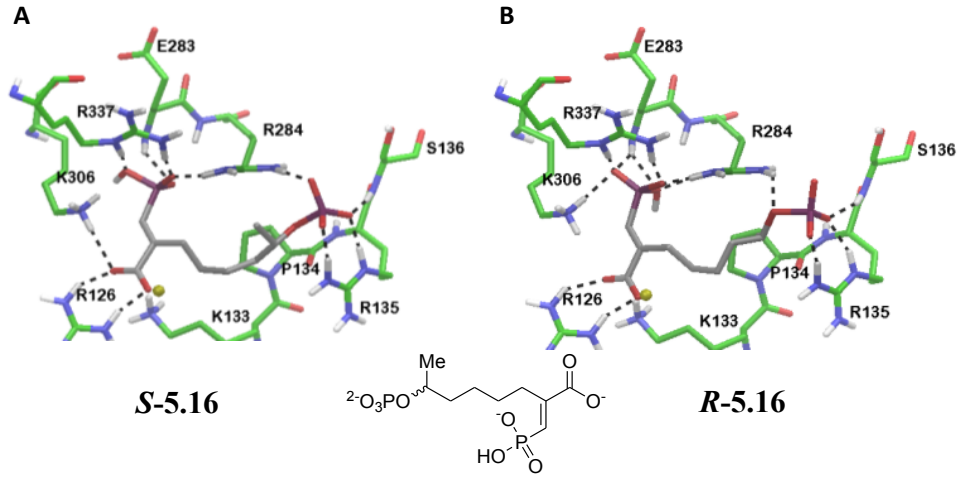


Figure 5. 18. Best poses from docking the *S*-isomer (A) and *R*-isomer (B) of structure 5.16 into the intermediate-adapted receptor conformation of *Mtu*DAH7PS.

Structure **5.16** is a geometric isomer of **5.13** (Figure 5. 18), with the *trans* geometry at the PEP-mimicking portion. Due to the *trans* geometry being a better mimic of the PEP molecule, this structure is predicted to bind deep in the active site pocket, as a result, the added methyl group at the E4P end is positioned in the hydrophobic environment of the active site. Unlike in structures **5.13**, **5.14** and **5.15**, where the methyl group is predicted to point towards solvent exposed space, here in **5.16**, the added methyl group provides extra hydrophobic contacts. This suggests that **5.16** may have higher binding affinity than its *cis* geometry analogues.

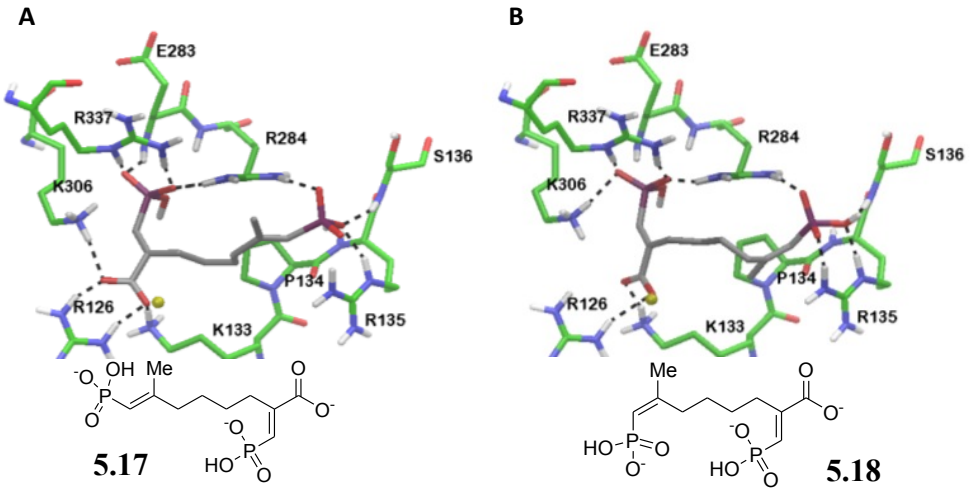


Figure 5. 19. Best poses from docking structure 5.17 (A) and 5.18 (B) into the intermediate-adapted receptor conformation of *Mtu*DAH7PS.

Structures **5.17** and **5.18** contain the same moiety at the PEP-mimicking portion as in **5.16**, but the E4P-mimicking end contains a double bond to restrict the positioning of the added methyl group (Figure 5. 19). Modelling suggests that by adding the double bond at the E4P-mimicking end, the steric clashes are reduced in **5.17** compared to **5.16**. In **5.18**, the added methyl group is forced to point inwards to the active site pocket and results in very favourable hydrophobic interactions in the modelling result. However, the buried methyl group also resulted in close contacts with the active site residues, and produces unfavourable van der Waals interactions.

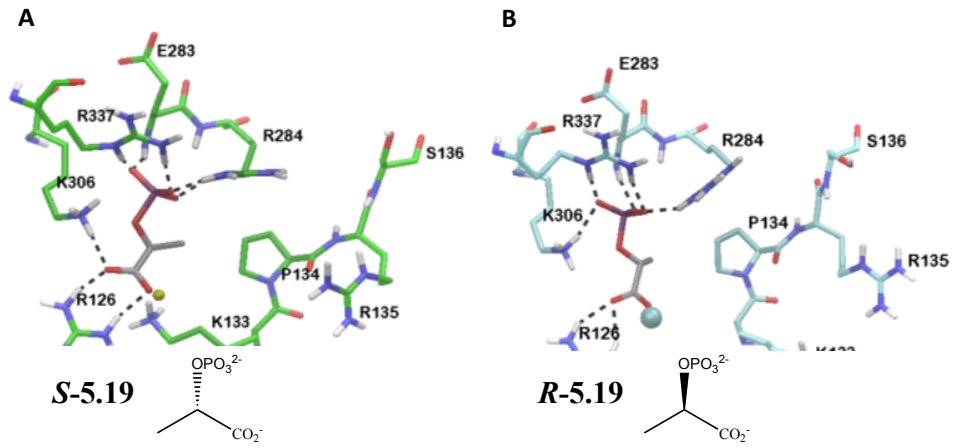


Figure 5. 20. Best poses from docking the *S*-isomer (A) of structure 5.19 into the intermediate-adapted receptor and the *R*-isomer in the original active site conformation (PDB 2B7O) of *Mtu*DAH7PS.

Structure **5.19** is a PEP mimic but with the tetrahedral geometry at C2 as in the intermediate. Modelling results show that in the intermediate adapted receptor grid, only the *S*-isomer of **5.19** can fit in the active site with retained orientation of phosphate and carboxylate groups as that of PEP found in the crystal structure. The *R*-isomer can fit with the retained orientation in only the receptor grid generated from the original crystal structure (2B7O). When both isomers retain the PEP orientations of phosphate and carboxylate moieties upon binding to the active site, due to the difference of absolute configurations at C2, the C3 methyl group will point to different directions. In the intermediate adapted receptor, due to the movement of the Lys133 side chain, the space behind the metal ion is more crowded than the original conformation in the crystal structure, in which Lys133 side chain is pointing away from the active site metal ion (Figure 5. 20). Therefore, the intermediate adapted receptor would select for binding poses with the methyl group pointing away from the Lys133 side chain, as observed for the *S*-isomer. While the receptor generated from the original crystal structure is more "open" and can accommodate both isomers with the same orientation. This selectivity does not predict that the *S*-isomer is preferred than the *R*-isomer, since the active site of *Mtu*DAH7PS is flexible and dynamic in reality, and may interconvert between the two or more conformations. Therefore, both isomers of **5.19** are predicted to be capable of binding to *Mtu*DAH7PS and may do so with similar affinity.

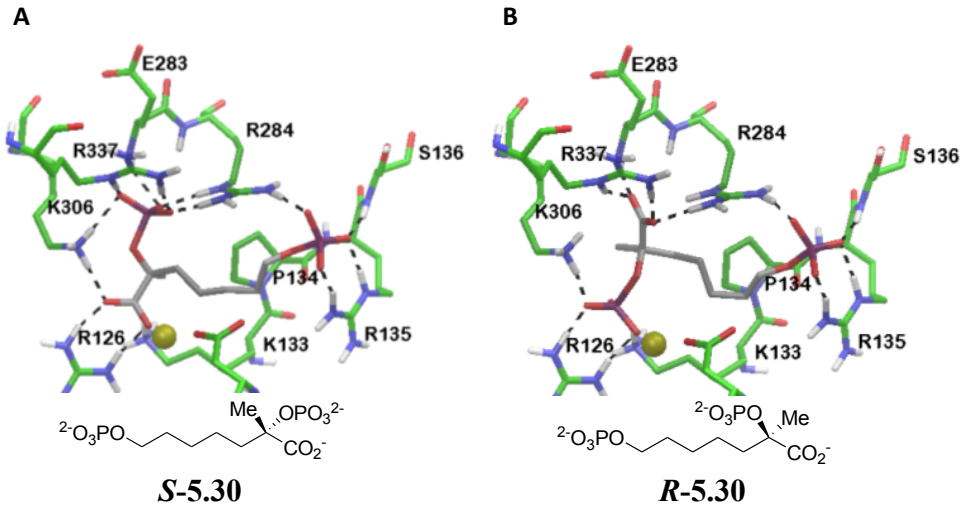
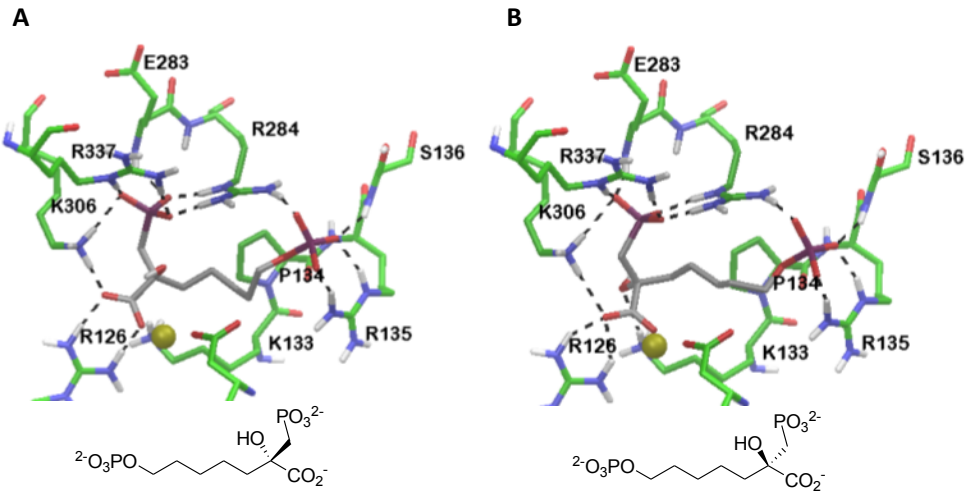


Figure 5. 21. Best poses from docking the *S*-isomer (A) and *R*-isomer (B) of structure

5.30 into the intermediate-adapted receptor conformation of *Mtu*DAH7PS.

Structure **5.30** is an analogue to **5.1**, but with an added methyl group at the C2 centre. The *R*-isomer of **5.30** does not model in the active site with the same orientation of the PEP-mimicking portion as that observed in the modelling of the tetrahedral intermediate (Figure 5. 21). This is because that if the PEP-mimicking portion is held in the correct orientation, the methyl group at C2 would point towards the metal ion in the active site, which may cause steric clashes and form unfavourable interactions. However, the *S*-isomer of **5.30** can bind in the same fashion as the intermediate molecule, with the methyl group at C2 positioned in the same binding pocket as the hydroxyl group in the linear intermediate. Consequently this monomer does not have a favourable hydrophobic score from modelling, due to the methyl group occupying the pocket surrounded by hydrophilic residues (e.g. Trp 280 and Glu248).



***S*-5.31** ***R*-5.31**

Figure 5. 22. Best poses from docking the *S*-isomer (A) and *R*-isomer (B) of structure 5.31 into the intermediate-adapted receptor conformation of *Mtu*DAH7PS.

The *R*-isomer of **5.31** is predicted to bind in similar mode as the linear intermediate (Figure 5. 22). Due to replacement of the oxygen atom with a methylene group in the phosphate moiety of the PEP-mimicking portion, the stereochemical configuration at C2 is reversed. This *R*-isomer of **5.31** has the same geometry as the *S*-isomer of the intermediate molecule. The *S*-isomer

of **5.31** can also fit in the active site in the retained PEP orientation, but does not have many favourable van der Waals interactions, due to steric clashes between the hydroxyl group at C2 carbon and residue His269 of the active site.

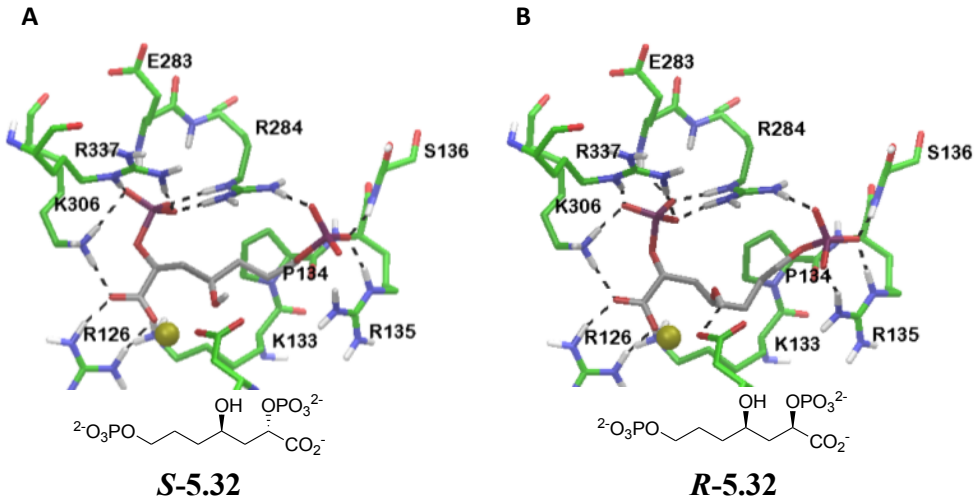


Figure 5. 23. Best poses from docking the *S*-isomer (A) and *R*-isomer (B) of structure 5.32 into the intermediate-adapted receptor conformation of *Mtu*DAH7PS.

The *R*-isomer of **5.32** binds with the retained orientation of the PEP-mimicking portion and in similar fashion to the intermediate molecule (Figure 5. 23). The hydroxyl group on C4 of *R*-isomer of **5.32** is 2.7 Å away from metal ion, suggests that it may be able to form coordination with the active site metal ion upon binding. The binding mode of the *S*-isomer is predicted to be quite similar to the *R*-isomer, with an extra hydrogen bond forming between the hydroxyl group at C4 and side chain of residue Asp441.

5.3 Docking studies for *E. coli* DAH7PS

During the course of the aldol reaction between PEP and E4P catalysed by DAH7PS, there are two intermediate species, the phosphohemiketal intermediate (**5**) with the tetrahedral geometry, and the oxocarbenium intermediate (**4**) with the trigonal planar geometry at the C2 carbon centre (Figure 5. 1). Whereas all intermediate inhibitors discussed above are

mimicking the tetrahedral intermediate **5**, another series of inhibitors based on the oxocarbenium intermediate **4** was designed and synthesised, and tested in *E. coli* DAH7PS (all experimental work mentioned in this section was conducted by Dr. Scott R. Walker). A molecular modelling study was therefore conducted for this series of inhibitors to explain the observed potencies .

E. coli DAH7PS is a member of the Type I family, and the crystal structures of both *E. coli* DAH7PS and *Mtu*DAH7PS show that although both enzymes adopt basic TIM barrel fold, the additional structural features responsible for allosteric regulation of the enzyme are quite different. However, the active site residues are well conserved due to the same catalytic function, and the interactions between bound substrates and the active site residues are retained in DAH7PS from both species (Figure 5. 24).

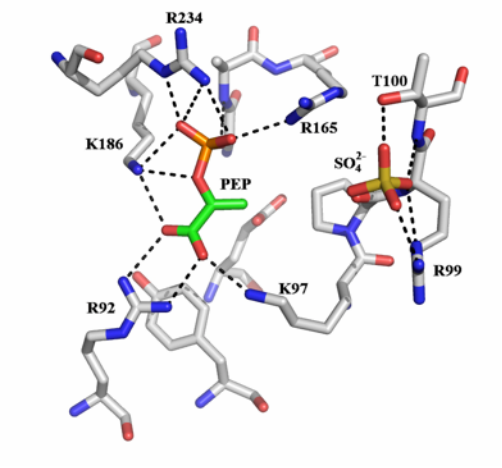


Figure 5. 24. The active site of phenylalanine-regulated DAH7PS from *E. coli* (PDB 1N8F) (*59*). A sulfate ion bound the active site interacts with Thr100 and Arg99, which is indicative of the E4P phosphate position when it is bound in the active site. The enzyme is displayed with white carbons and the active site bound PEP is shown with green carbon atoms. Hydrogen bonds are displayed as black dashed lines.

Previously, a series of PEP analogues was synthesised and investigated as inhibitors to examine the ability of inhibitors to target the PEP site of the DAH7PS from *E. coli* (*86*). Compounds bearing both carboxylate and phosphonate or phosphate groups separated by a similar distance to those found in PEP were found to be effective competitive inhibitors with respect to PEP. The best inhibitors of the enzyme were compounds that included a

trigonal centre adjacent to the carboxylate moiety, such as allylic phosphonate **7** and vinyl phosphonate **8** (Figure 5. 25). These compounds were considered to be mimicking the intermediate oxocarbenium ion **4**.

Previous single site inhibitors were then elaborated to include an appropriately spaced distal phosphate moiety, positioned to be capable of interacting with the binding pocket for the corresponding phosphate moiety of co-substrate E4P (Figure 5. 24). This alteration was made in order to allow the inhibitors to bind to both PEP and E4P binding sites simultaneously, utilising the principle of multivalency (*87*).

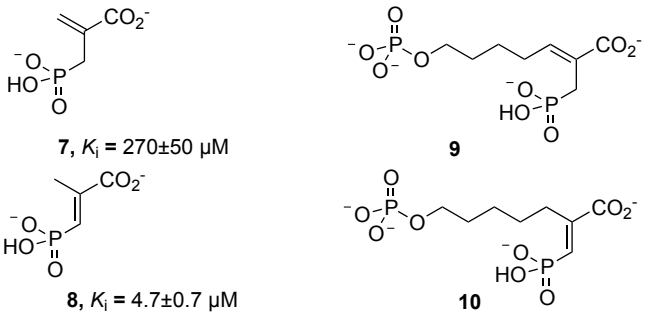


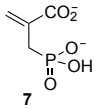
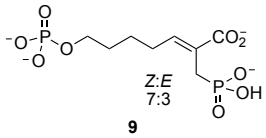
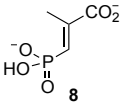
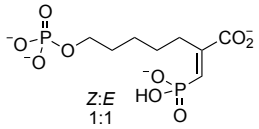
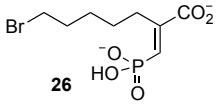
Figure 5. 25. Parent PEP-competitive inhibitors allylic phosphonte 7 and vinyl phosphonate 8, and their corresponding extended inhibitors 9 and 10.

In order to understand the binding modes and contributions from active site residues for ligand binding, and hence explain the trends observed from experimental measurements, the designed extended compounds were modelled into the active site of *E. coli* DAH7PS using induced fit docking protocol. Similar to the induced fit docking set up used in *Mtu*DAH7PS (discussed above), here the active site in chain A of the *E. coli* DAH7PS crystal structure (1NF8) was used as the receptor. The centre of the grid was defined as the centroid of residues 92, 97, 98, 99, 165, 186, 268 and 234. For the initial docking, side chains of residues 268, 186, and 97 were removed because they were blocking access to the active site, and the van der Waals radii of the docked inhibitors were scaled to 0.8. The 20 best poses of initial docking were kept. All residues of the enzyme within 5 Å of the docked inhibitor poses were refined to optimise interactions. At this stage the side chains of residues 268, 186, 97 and 99 were also included. The inhibitors

were then re-docked, to the top 20 newly generated protein structures if the energy was within 30 kcal/mol of the best protein structure.

The ability of the extended compounds to inhibit *E. coli* DAH7PS was determined by kinetic assay (
Table 5. **2**). The diastereomeric mixture of extended allylic phosphonate **9** was found to give competitive inhibition with respect to PEP, with an inhibition constant K_i of 154±20 μM. The inhibition constant is an improvement over that found for the parent PEP-like inhibitor **7**, demonstrating the advantages of a dual-site approach for DAH7PS inhibition. To understand the binding modes and contributions of each isomer in the mixture to the inhibition, the two isomers were modelled into the *E. coli* DAH7PS active site separately using induced fit docking (Figure 5. 26). The (*Z*)-isomer of allylic phosphonate **9** showed a reversed binding mode, with the PEP portion of the molecule binding in the E4P site and the E4P phosphate group binding in the PEP site. This suggests that the (*Z*)-geometry of this compound at the double bond does not fit well with the shape of the active site binding pocket. As a result, the PEP portion of the compound was placed in the more accommodating E4P site. However, the (*E*)-isomer of phosphonate **9** showed the expected binding mode, with the PEP-like and E4P-phosphate like portions placed in the designated binding site. The carboxylate group on the PEP side of the (*E*)-isomer shows interactions with Arg92 and Lys186; the phosphonate group on the PEP side interacts with Lys186, Arg234 and Arg165. These interactions are equivalent to those observed in the original crystal structure between the bound PEP molecule and the DAH7PS active site (Figure 5. 24). The E4P-like phosphate group of the (*E*)-**9** forms interactions with residues Arg165, Lys273, Arg99 and Thr100. The (*E*)-isomer of allylic phosphonate **9** is a better mimic of the reaction intermediate **4** and therefore is likely to largely account for the inhibition of *E. coli* DAH7PS observed for the isomeric mixture.

Table 5. 2. Inhibition constants of the parent compounds and the extended compounds.

Parent Compound	K_i (μM) Parent compound	Extended compounds	K_i (μM) Extended compound
 7	270±50	 9	154±20
 8	4.7±0.7	 10	5.3±0.6
		 26	3.6±0.4

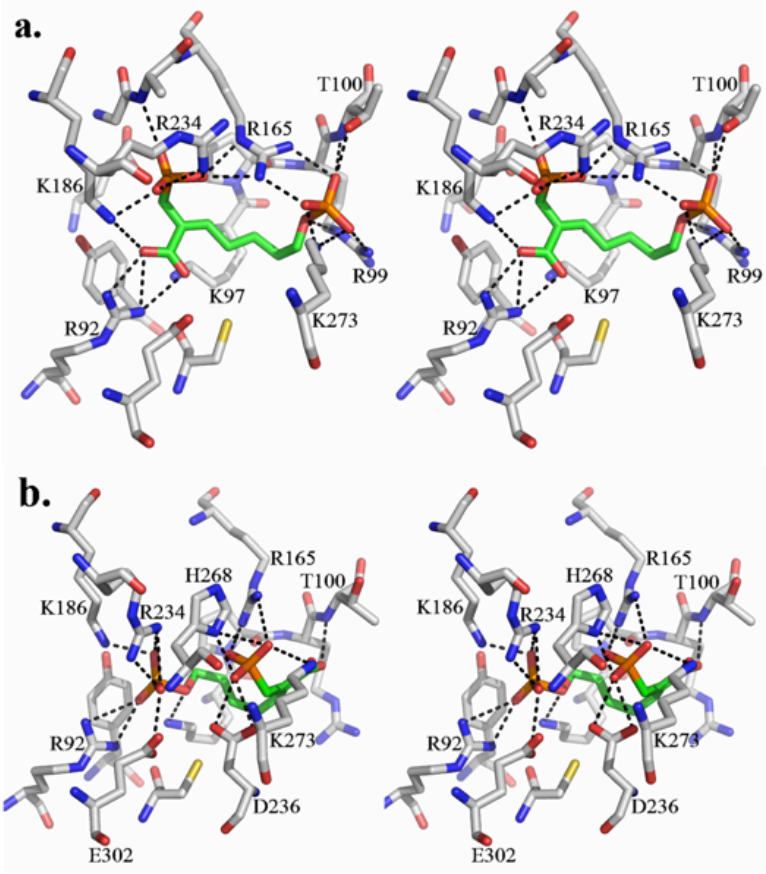


Figure 5. 26. Poses (in stereoview) from induced fit docking calculations for compounds (a) (*E*)-9 and (b) (*Z*)-9. Atoms on the enzyme active site are displayed with white carbons and the inhibitor molecules are displayed with green carbons. Hydrogen bonds are showed as black dashed lines.

The vinyl phosphonate **10** was obtained as a 1:1 mixture of (*Z*)- and (*E*)-isomers. This mixture was determined to competitively inhibit *E. coli* DAH7PS with respect to PEP. The vinyl phosphate **10** was a potent inhibitor of DAH7PS, with an inhibition constant K_i of 5.3±0.6 μM. This inhibition constant is similar to the unelaborated parent compound **8**, however this compound consisted solely of the desired (*E*)-isomer, and the corresponding (*Z*)-isomer is a much weaker inhibitor of DAH7PS (*86*). The inseparable mixture of isomers in **10** may obscure the anticipated increase in binding strength expected for the originally designed (*E*)-isomer of the vinyl phosphonate **10**. To understand the binding mode and potency of each isomer in the compound **10** mixture, the (*E*)- and (*Z*)-isomers of compound **10** were modelled into the active site of *E. coli* DAH7PS separately (Figure 5. 27A and B). Both isomers showed the expected binding mode with the PEP and E4P portions positioned in the desired sites. The (*E*)-isomer fits well in the active site (Figure 5. 27A), with the PEP phosphate-like phosphonate group interacting with Arg234 and Arg165. The side chains of Lys186 and Arg92 interact with the carboxylate group. The E4P-like phosphate group forms interactions with Arg165, Arg99 and Thr100. In the (*Z*)-isomer model, the PEP-like portion does not fit deep in the PEP site, presumably due to the change in relative geometry of the phosphonate and carboxylate groups (Figure 5. 27B). As a result, interactions between the inhibitor carboxylate group and Arg92 are lost in compound (*Z*)-**10**; a carboxylate-Arg92 binding interaction is observed for PEP in the original crystal structure (Figure 5. 24) (*59*). Therefore, the modelling results suggest that the (*E*)-isomer of compound **10** binds to the enzyme active site better than the corresponding (*Z*)-isomer.

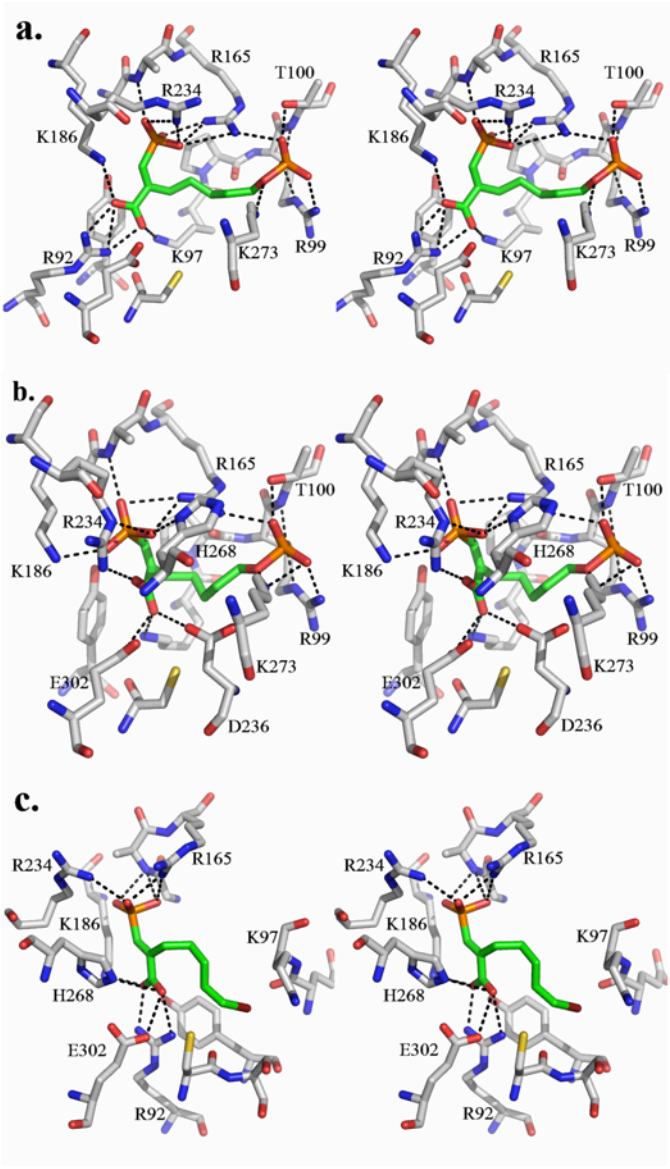


Figure 5. 27. Poses (in stereoview) from induced fit docking calculations for compounds (a) (*E*)-10, (b) (*Z*)-10, and (c) 26. Atoms on the enzyme active site are displayed with white carbons and the inhibitor molecules are displayed with green carbons. The Hydrogen bonds are showed as black dashed lines.

The bromide **26**, a side product from synthesis of compound **10**, was also evaluated as an inhibitor of *E. coli* DAH7PS. It too proved to be an inhibitor of *E. coli* DAH7PS, with a K_i of 3.6±0.4 µM. The potency of bromide **26** was initially surprising, since it lacks the terminal phosphate group designed to enhance the binding of these compounds to DAH7PS. Modelling of this compound showed that the PEP portion of bromide **26** (Figure 5. 27C) is in the same position as the PEP molecule found in the crystal structure (Figure 5. 24). As expected, the bromide compound **26** is not picking up distal interactions at the E4P-phosphate binding site. However, the K_i value for this

compound is not significantly different from that of the extended vinyl phosphonate compound **10**, or the parent compound **8**. The similar K_i values may be due to the bromide compound **26** consisting of only the PEP-like (*E*)-isomer of the vinyl phosphonate group, whereas compound **10** also contains the interfering (*Z*)-isomer. The presence of the hydrocarbon tail therefore does not appear to impede the ability of the PEP-mimicking portion of bromide **26** to occupy the PEP binding site of *E. coli* DAH7PS. From the modelling results, the side chain of Lys97 moved away from the PEP site which creates a binding pocket for the bromide. This added hydrophobic interaction between the hydrocarbon linker of bromide **26** and the enzyme active site may also contribute to the potency of this inhibitor.

5.4 Summary

In this chapter, the technique of molecular modelling was employed to validate the methodology and to explain observations from existing experimental data. Having validated the methodology it was used to predict potency of proposed structures and hence guide the future design of potential drug-like compounds for DAH7PS.

Although many aspects of the reaction mechanism catalysed by DAH7PS are known to date, the stereochemistry for the water attack remains unclear (Figure 5. 1). Induced fit docking was conducted for both diastereoisomers of the tetrahedral intermediate molecule (**5**), in order to identify preferred stereochemistry for the water attack. The results showed the active site of *Mtu*DAH7PS exhibits a preference for the *S*-isomer of the linear intermediate, which suggests that it is preferred that water attacks from the *re* face of the oxocarbenium species (**4**). Important active site residues involved in substrate binding were also identified from the modelling results, and they include Arg126, Lys306, Arg284, Arg337, Arg135, Ser136, Trp280, Glu284 and Lys133.

Induced fit docking of the intermediate mimicking compound **5.1** predicted very similar binding modes to that of the linear intermediate molecule (**5**), and interestingly the predicted binding mode closely resembles the binding mode observed in the crystal structure of *Mtu*DAH7PS co-crystallised with racemic mixture of **5.1**. This demonstrates that the method of induced fit docking is suitable for predicting binding modes and potencies of the series of intermediate mimicking inhibitors. The predicted binding modes and potencies of the series of intermediate mimicking compounds (discussed in section 5.2.5.1) that have not yet been synthesised will need to be verified by experiments.

Chapter 6. Investigating the molecular basis of synergistic allostery of *Mtu*DAH7PS

6.1 Introduction

6.1.1 Allosteric regulation

Allosteric regulation of protein function is critical for the control of metabolism. Allostery is the process by which remote sites of system are energetically coupled to give functional change. This mechanism of regulation is widely observed in biological systems and has been recognised for many years (*88-94*). The most prominent examples of allostery have featured proteins where ligand binding was associated with significant conformational changes, such as haemoglobin and aspartate transcarbamoylase (*89, 90*). More recently, this view of allostery has been refined, and allosteric mechanisms including far more subtle changes in protein conformational states have been recognised, contributing to the growing understanding of the importance of protein dynamic motion to the catalytic function of enzymes (*95-106*). The current understanding of allostery is based on the concept of "population shift" of the protein ensembles, which is mediated by alteration of the two thermodynamic components, enthalpy (gross conformational change) and entropy (change in protein dynamics), and systems at both extremes have been examined. For example, MD simulations have been used to study the synergistic regulation of tryptophan synthase complex, which is a bifunctional tetrameric αββα enzyme complex (*107*). This complex illustrates cooperative motions and conformational changes of α- and β- subunits by tightly controlled allosteric interactions. Studies using MD simulations reveal that, in the ligand-free state, both α- and β- subunits of the enzyme samples a range of conformations including open and partially closed conformations, and that binding of ligands shifts the conformational ensemble towards the fully closed conformations. In contrast, the binding of two cAMP ligands to the

catabolite activator protein (CAP) system results in negative cooperativity with no average conformational changes (*108*). Experimental studies have shown that binding of the first cAMP ligand to CAP does not result in any conformational changes of the ligand binding site in the other subunit; however, the dynamics of the system is altered so that the binding of the second cAMP ligand is associated with a higher entropic penalty, thus reducing the binding affinity of the second cAMP ligand by two orders of magnitude (*108*).

Allosteric regulation of *Mtu*DAH7PS is proposed to be mainly entropy driven, as structural changes on binding the allosteric regulators are subtle (see Chapter 4) (*76*). Unlike other DAH7PSs, single amino acids do not alter enzymic activity, yet binary combinations of aromatic amino acids give rise to a synergistic response. Combinations of Trp and Phe, and to a lesser extent Trp and Tyr, result in significant inhibition of the enzyme, providing a sophisticated mechanism for the control of a branched biosynthetic pathway. Moreover, in the presence of inhibitory concentrations of Trp and Phe or Trp and Tyr, significant homotropic cooperativity was observed in enzyme activity with respect to the concentration of substrate E4P, providing another level of sophistication to the communication between inhibitor and substrate binding sites.

In this chapter, the communication among the allosteric binding sites and the active sites of *Mtu*DAH7PS will be investigated using a combination of MD simulations and bioinformatics techniques (statistical coupling analysis). These studies illuminate the central role that subtle changes in protein dynamics play in the unusual synergistic inhibition of this enzyme.

6.1.2 Application of the statistical coupling analysis

One of the computational methods used to study allostery is the sequence based statistical coupling analysis (SCA) (*109*). This analysis has been used to identify amino acid networks that are responsible for communication within proteins. It uses the evolutionary data of a protein family to measure

energetic coupling between positions on a multiple sequence alignment (MSA). This analysis is based on the view that evolution of a protein fold is the result of a large scale random mutagenesis, with selection constraints imposed by function.

The SCA method is based on two hypotheses. First is that lack of evolutionary constraint at one position should cause the distribution of observed amino acid at that position in the MSA to approach their mean abundance in all proteins, and deviance from the mean values should quantitatively represent conservation. The second hypothesis is that, the functional coupling of two residues, even when they are distantly located in the structure, should mutually constrain evolution at the two positions, and these should be represented in the statistical coupling of the underlying amino acid distributions.

The capability of the SCA method to predict coupled amino acids distant from the active site has attracted huge attention in the research community since its introduction (the original paper introducing the SCA method (*109*) published in 1999 has been cited 424 times so far according to data on Web of Science ^{SM}), because residues involved in catalysis or allostery but not in the near vicinity to the active site are generally hard to predict by other means of experimental methods. The SCA method has been used to study co-evolved amino acids in various protein families, including PDZ domains, G-protein coupled receptors, M.Hhal, iLBPs, aspartic proteinases, $(\beta/\alpha)_8$ barrel proteins, and superoxide dismutases (*110-116*).

New techniques have been developed which combine the SCA with other computation methods, such as MD simulations. The SCA method was used in combination with MD simulations to identify amino acid pairs essential for catalysis in the M.Hal family (*111*). In this work, a hybrid method, denoted SCA·MD, which weights the correlated motion from MD simulations with evolutionary conservation from SCA, was developed. The hybrid SCA·MD matrix was obtained by multiplying elements in the SCA correlation matrix with the ones in the MD correlated motion matrix. In the

SCA part, the correlation values were not calculated for all positions on the MSA. Positions on the termini were omitted. In the MD part, the backbone α carbon movement was considered. The hybrid matrix predicted several pairs of amino acids with high correlated or anti-correlated motions. Some of the predictions were tested by mutagenesis experiments, in which one of the amino acids in the pair were mutated into Ala, and the changes in k_{cat} were measured as indication of the mutation effect. Experimental results showed that mutation of amino acids predicted from the SCA·MD method produced large effect on rate of catalysis, while mutation of residues adjacent to the predicted residues did not produce the same effect. The SCA technique is more powerful when combined with MD simulations, for identification of subset of co-evolved residues which show coupled motions within the protein. The results from SCA and MD simulations are generally visualised with crystallography and tested by mutagenesis experiments.

In this chapter, the molecular basis of the communication mechanism of synergistic allostery in *Mtu*DAH7PS was examined. MD simulations were conducted for ligand-free and ligand-bound *Mtu*DAH7PS complexes, in order to examine the changes in dynamic fluctuations of individual residues caused by allosteric ligand binding. The SCA method was applied for type II DAH7PS family, of which *Mtu*DAH7PS is a member, in order to identify the co-evolved residues within this protein family. Finally the SCA method was combined with results from the MD simulations using the hybrid SCA·MD approach, to identify residues that are critical for the communication between the allosteric sites and the active site in *Mtu*DAH7PS.

6.2 MD simulations for *Mtu*DAH7PS in complex with different ligands

6.2.1 ITC measurements of ligand binding affinity.

The unusual synergistic inhibition of *Mtu*DAH7PS is only observable in the presence of both Trp and Phe. In order to probe the binding of single ligands, isothermal titration calorimetry (ITC) was employed (ITC experiments were conducted by Dr. Richard Hutton in our research group). These results are reported here as they are important to the interpretation of the computational studies described in this chapter. Titrations of the allosteric inhibitor Phe yielded a dissociation constant (K_d) of 21.2 μM. Binding experiments with the alternative allosteric inhibitor, Trp, yielded a K_d value of 4.7 μM, indicating clearly that Trp binds to *Mtu*DAH7PS with higher affinity than Phe (
Table 6. **1**).

Table 6. 1. Results from ITC binding experiments.

Binding experiment (298 K)	*K_d (μM)*	*Stoichiometry*
Phe titration	21.2 ± 1.0	-
Phe titration (20 *μ*M Trp background)	5.0 ± 0.1	3.98 ± 0.02
Trp titration	4.7 ± 0.1	4.02 ± 0.03
Trp titration (20 *μ*M Phe background)	1.08 ± 0.03	3.96 ± 0.01

To examine the influence that the presence of one ligand has on the affinity of the protein for the other ligand, ITC experiments were performed using a background of one of the allosteric inhibitors in the sample cell whilst titrating in the other. Titration of Phe onto a Trp background yielded a reduced K_d of 5.0 μM. Inversely, titrating in Trp to a Phe background yields a reduced K_d of 1.08 μM. These experiments clearly show that both binding sites do not just communicate their occupancy to the active site to influence catalysis, but that the Phe and Trp sites exchange information on their occupancy directly.

6.2.2 MD simulations for Mtu*DAH7PS complexes*

MD simulations were conducted to examine the change of dynamic properties of the *Mtu*DAH7PS tetramer upon binding of different ligands. MD simulations (20 ns) were set up for four systems of the *Mtu*DAH7PS tetramer: ligand-free, Trp+Phe, Trp-only, and Phe-only bound.

The MD simulations were carried out with NAMD running on the BlueFern supercomputer at the University of Canterbury (*36*). Four crystal structures of *Mtu*DAH7PS, including ligand-free (PDB 3NV8), Trp+Phe co-crystallised (PDB 3RZI), Trp-only (PDB 3NUE) and Phe-only (PDB 3NUD) bound, were used in the MD simulations. Missing residues in the three ligand-bound structures of *Mtu*DAH7PS were built in with COOT in accordance with the electron density map before MD simulations (*117*), so that the dynamic properties of the disordered loops can also be examined. The Phe-only soaked structure (3NUD) (*76*) contains Phe molecules in primary and secondary Phe sites and Trp binding sites. For MD simulations, only Phe molecules in primary Phe binding sites were kept, in order to investigate the dynamic behaviour of the enzyme when the ligand molecules were only at their preferred sites (the preferred binding sites for Phe molecules are at the dimer interface, this has been validated by both the co-crystallised Trp+Phe structure (Chapter 4) and mutagenesis experiments (*76*)). The tetrameric form of each complex was generated from the dimer in the asymmetric unit based on the 2-fold crystallographic symmetry operation. Each of the four complete tetramer structures was solvated with explicit TIP3 water molecules in a box in VMD (*118*) ionised by adding Na^+ and Cl^- ions to balance the net charge of the water box. The ions were added with a minimum distance of 5 Å from the molecule and each other. MD simulations were conducted with the CHARMM force field parameter specifications at a constant temperature and pressure (310 K, 1 atm) (*41*). The cut-off distance for van der Waals interactions was set to 12 Å. In each simulation, the system was first minimised for 1000 steps followed by 20 ns of dynamics simulation conducted with 2 fs time steps. The trajectory was written out at 100 ps time intervals, and a total of 200 frames were obtained from each simulation (20

ns).

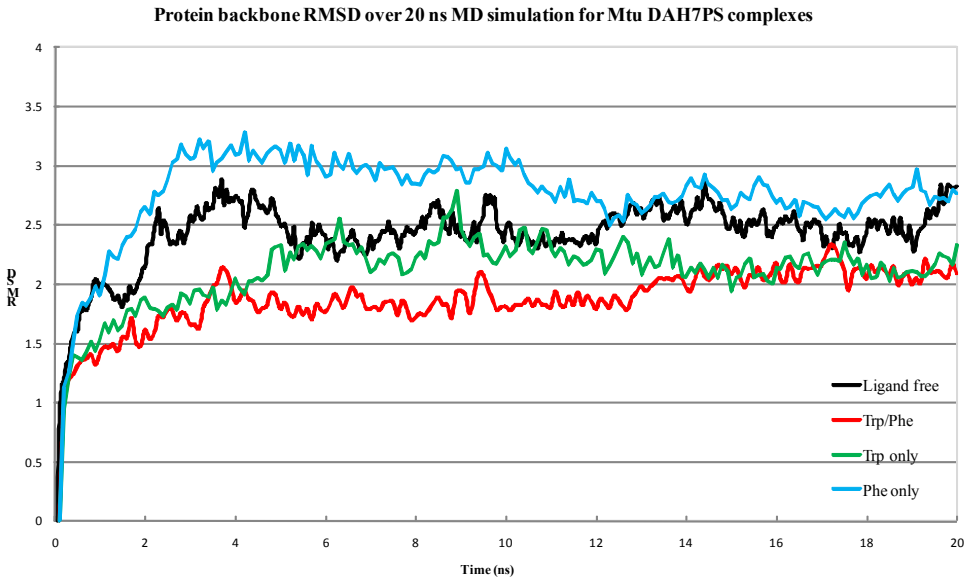


Figure 6. 1. Backbone RMSD of *Mtu*DAH7PS in complex with different ligands during 20 ns MD simulations. The RMSD values were calculated using the first frame of the trajectory as the reference structure.

Protein backbone RMSD values (Figure 6. 1) indicate that the backbone RMSD values of the protein molecule in all four simulations become stable after ~2.3 ns, indicating that all for simulations equilibrated after ~2.3 ns. Within the equilibrated time period (2.3 ns to 20 ns), all four complexes showed small fluctuations in backbone conformation. The average conformations of the Trp+Phe bound, Trp-only and Phe-only bound systems during equilibrated time period match to that of the ligand-free system with RMSD values of 1.562 Å (matching 1751 C^{α} atoms), 1.375 Å (matching 1714 C^{α} atoms) and 1.326 Å (matching 1659 C^{α} atoms), respectively. Considering the extent of movement within each system (~1.5 – 3 Å) during the MD simulations, these RMSD values are relatively small, and average conformations of the four systems differ only slightly, consistent with the absence of major conformational change in *Mtu*DAH7PS upon allosteric ligand binding observed crystallographically.

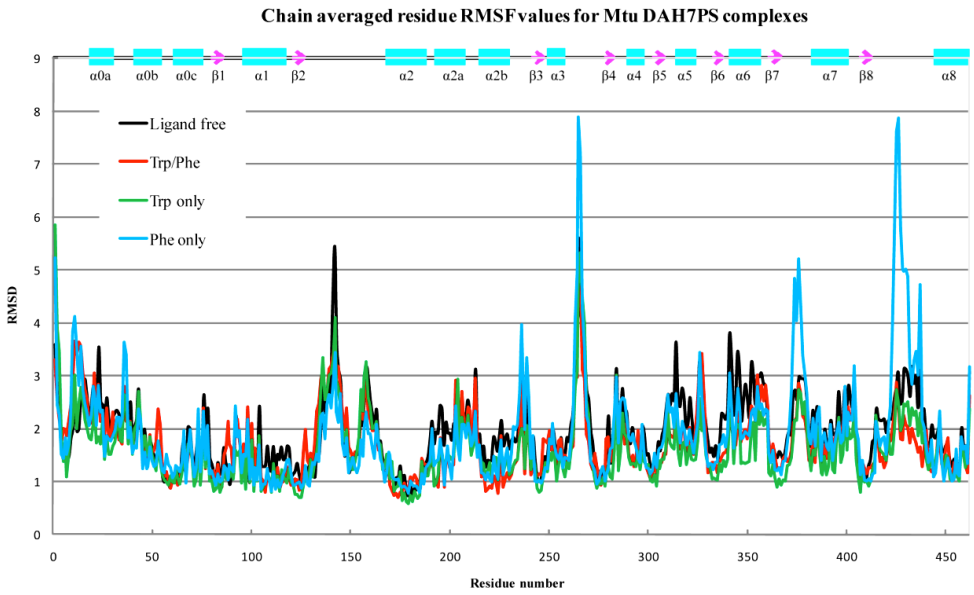


Figure 6. 2. Chain-averaged RMSF values residue-by-residue for the four *Mtu*DAH7PS complexes. The RMSF values were calculated using the trajectory frames collected during the equilibrated simulation period, and all non-hydrogen atoms in each residue are included in the calculation (including backbone and side chain).

In order to examine the changes in flexibility of specific regions of the structure upon ligand binding, the RMSF (root mean square fluctuation) values of individual residues of *Mtu*DAH7PS were calculated. The RMSF values were calculated for each amino acid using its average conformation during the MD simulation as the reference structure, and therefore these values provided an indication of individual residue conformational fluctuation during the MD simulations. For this calculation, trajectory frames collected from the equilibrated simulation period were used (after 2.3 ns), and all non-hydrogen atoms of each amino acid (backbone and side chain atoms) were included. During the MD simulations, the four identical monomers of the tetrameric *Mtu*DAH7PS can be considered as sampling slightly different regions of the same conformational space. Therefore, instead of comparing residue RMSF values of individual monomers, comparison of the RMSF values averaged over the four monomers was conducted to give a better indication of regions of flexibility (Figure 6. 2). As expected, for all four systems, the residues on the loop regions gave greater RMSF values (more flexible) and the residues on secondary structures such as helices and sheets showed relatively smaller RMSF values (less flexible). The differences in residue RMSF values (ΔRMSF) between the ligand-bound

systems and the ligand-free system were then calculated so that positive ΔRMSF values indicate greater flexibility and negative values indicate less flexibility in the corresponding region upon ligand binding (Figure 6. 3).

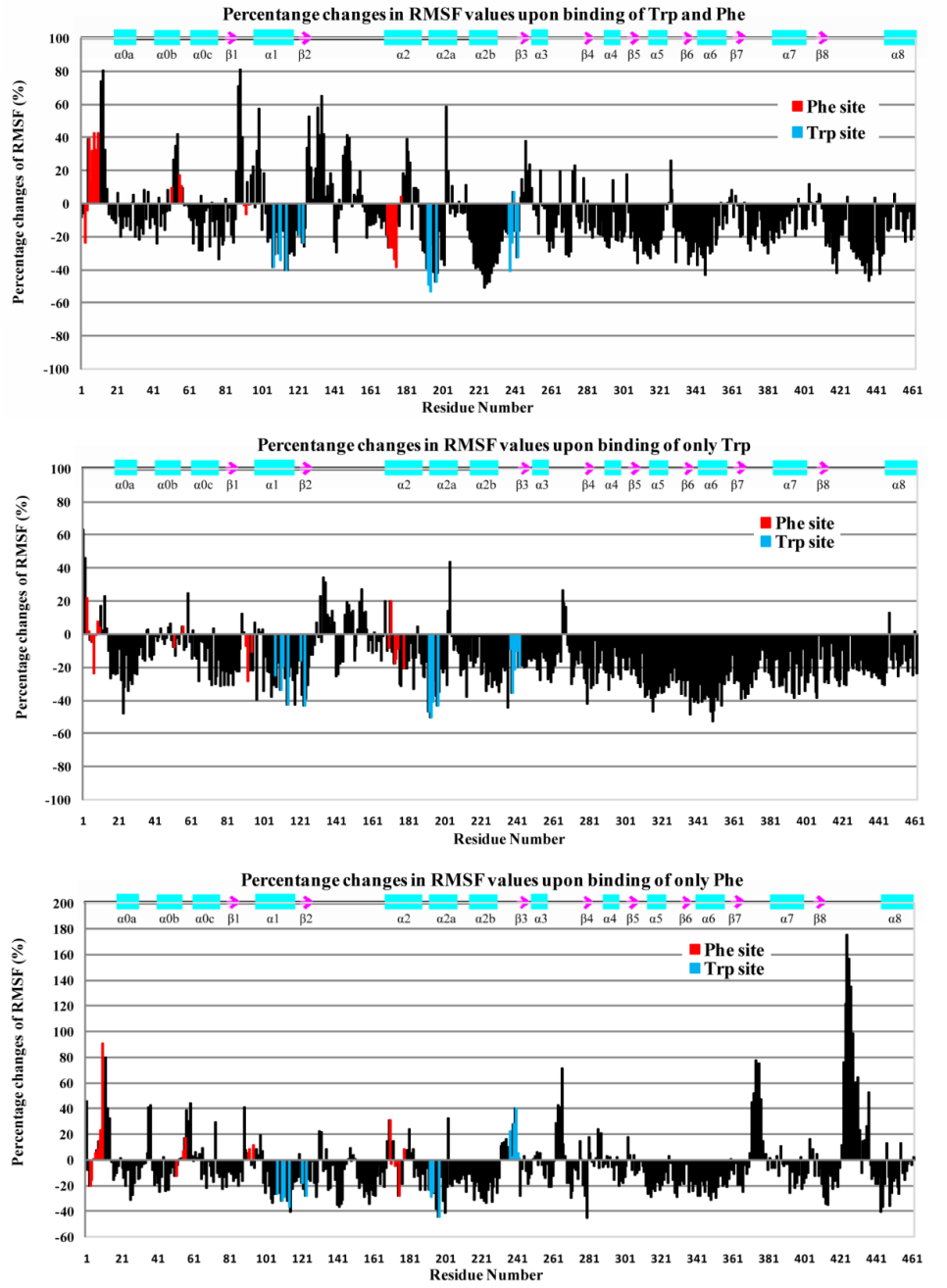


Figure 6. 3. Percentage changes in flexibility (ΔRMSF) with respect to wild type *Mtu*DAH7PS upon binding of different ligands to *Mtu*DAH7PS; positive values indicate increased flexibility in the corresponding region and negative values indicate reduced flexibility upon binding of (A) Trp+Phe, (B) Trp-only, and (C) Phe-only, as revealed by molecular dynamics simulations. Residues involved in Trp and Phe binding sites are coloured in cyan and red, respectively.

6.2.2.1 Change in flexibility upon binding of both Trp and Phe

Upon binding of both Trp and Phe in the allosteric sites (Figure 6. 5A), the tetramer interface becomes more rigid and the N-termini and part of the dimer interface become more flexible. Notably, the loop (β2-α2) containing residues 133-135 of the active site, becomes more flexible when both Trp and Phe are bound (ΔRMSF values are ~0.840 Å on average for residues 133-135). The alteration in regional flexibility is also confirmed by an analysis of the *B*-values (temperature factors) between the Trp+Phe co-crystal and ligand-free structure of *Mtu*DAH7PS (**Error! Reference source not found.**). The β2-α2 loop is responsible for the binding of the substrate E4P (*61*), (*62*) and has recently been shown to bind to the E4P-mimicking portion of an *Mtu*DAH7PS inhibitor designed to resemble the reaction intermediate between PEP and E4P (*85*). The increased flexibility of loop β2-α2 increases the entropy penalty upon binding of the substrate E4P, and thus disfavours E4P binding. The MD result is entirely consistent with experimental observations: kinetic analysis has earlier revealed that the strongly inhibitory effect of the combination of Trp and Phe is focused on a dramatically reduced affinity of the enzyme for E4P (*76*).

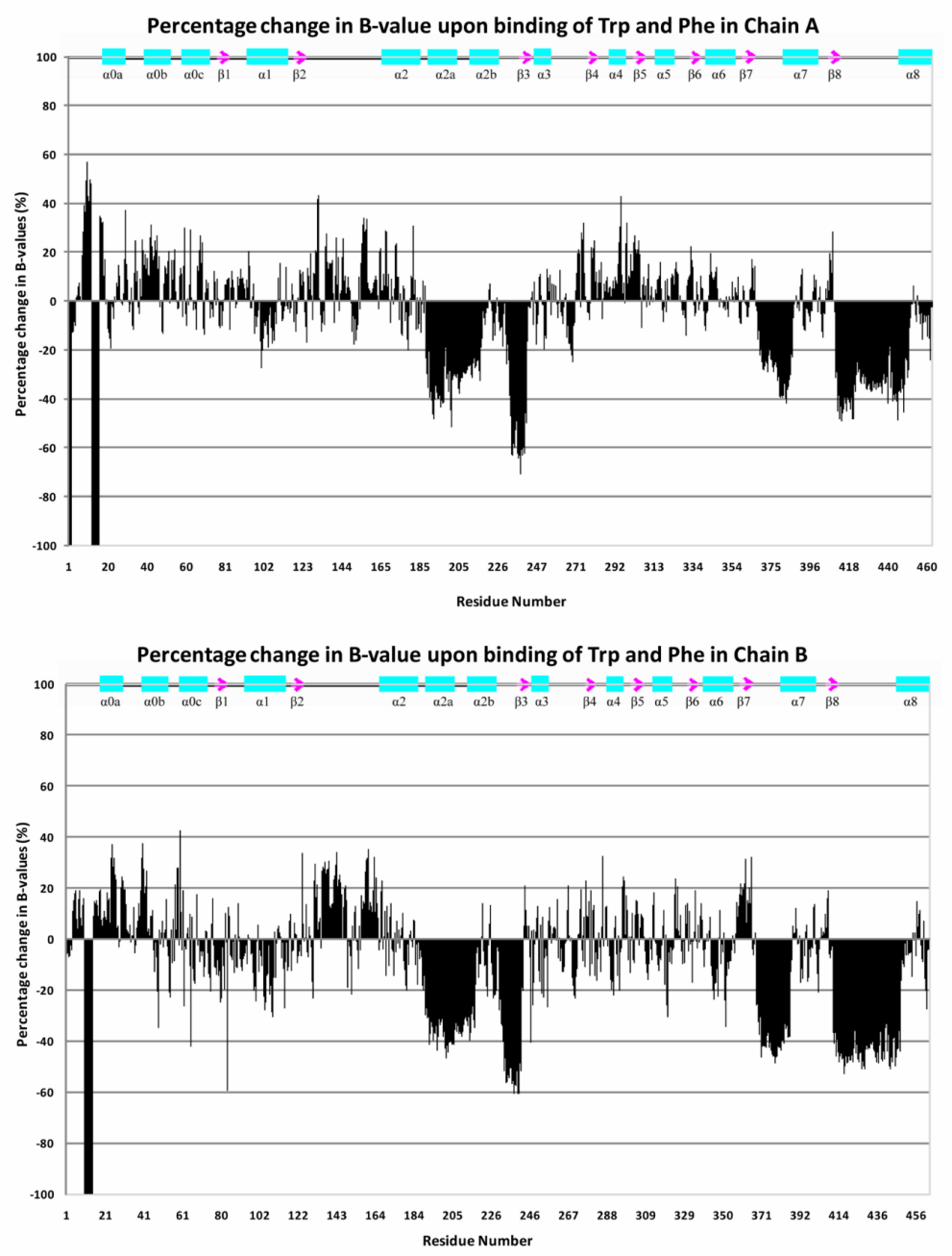


Figure 6. 4. Plot of percentage differences in *B*-value of CA and CB atoms in each residue between wild-type *Mtu*DAH7PS with and without Trp+Phe (where Trp+Phe was co-crystallised), comparison of *B*-value in Chain A is plotted in top frame, and that of Chain B is plotted in bottom frame. Positive values indicate increased *B*-values (flexibility) on binding Trp+Phe; Negative values indicate decreased *B*-values (flexibility on binding Trp+Phe, and -100% indicates residues missing in the Trp+Phe structure.

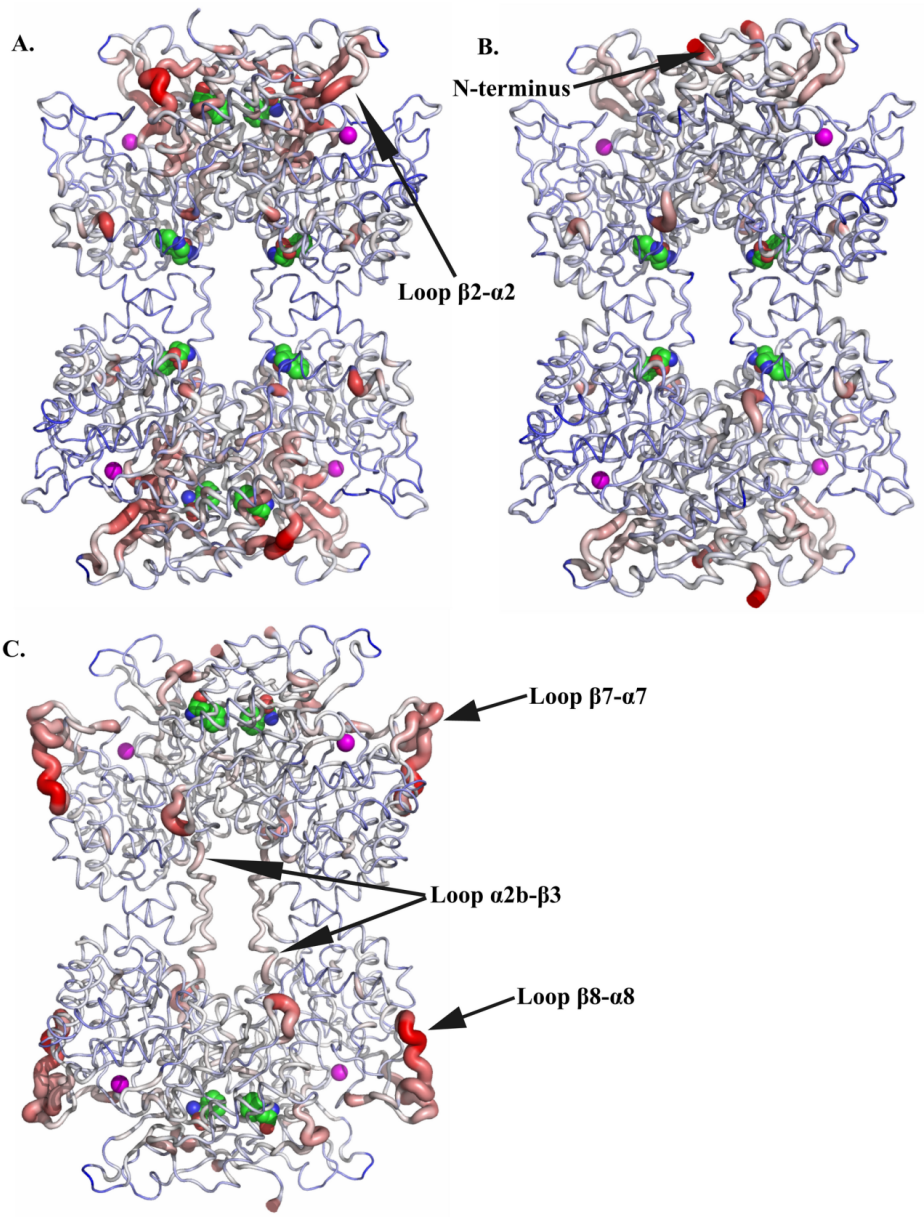


Figure 6. 5. Changes in residue flexibility (ΔRMSF) in *Mtu*DAH7PS upon binding of (A) both Trp and Phe, (B) Trp-only and (C) Phe-only, as revealed by molecular dynamics simulations. The changes in flexibility are displayed both by colour and tube width. The regions with increased flexibility are coloured with changing gradient of red, and thicker tube width; regions with decreased flexibility are coloured with changing gradient of blue, and thinner tube width.

6.2.2.2 Change in flexibility upon binding of single amino acids

One of the intriguing aspects of the unusual synergistic response to Trp and Phe binding in *Mtu*DAH7PS is that no alteration in catalytic activity of the enzyme on binding of a single ligand is observed, yet ITC results show clearly that the two sites are in mutual communication. Therefore, MD trajectories of Trp-only and Phe-only systems were analysed to investigate dynamic changes in the active site and the allosteric binding sites associated with single ligand binding.

Upon binding of Trp alone (Figure 6. 5B), the tetramer interface and Trp binding sites become more rigid, in a manner similar to that observed for the Trp+Phe system. However, surprisingly, the binding of Trp alone was associated with an increase in flexibility of the distant N-terminal region (residue 1-4, 9-10), which caps the Phe binding pocket. Other residues constituting the Phe binding sites, namely the β1-α1 loop (residues 91-94) and the α2 helices (residues 170-178) in the dimer interface become slightly less flexible upon Trp binding. These observations suggest that in the presence of Trp, the binding of Phe will be enhanced as most residues in the Phe binding sites become less flexible, decreasing the entropic cost of Phe binding. The more flexible capping N-terminal residues facilitate access to the Phe binding sites. The active site loop β2-α2 (residues 133-135) also showed a slight increase in flexibility (ΔRMSF is ~0.359 Å for residues 133-135 on average), but this increase is much smaller than that calculated when both Trp and Phe are bound (0.840 Å), indicating the most prominent effect on active site loop dynamics (hence inhibitory effect) is only achieved by binding of both Trp and Phe.

The binding of Phe alone (Figure 6. 5C) also resulted in slight increases in flexibility of the dimer interface and the N-terminal region. Loops β7-α7 and β8-α8 showed unusually large increases in flexibility in the Phe-only system, in conjunction with the fact that these loops were also missing in the original crystal structure due to disorder, it indicates that the highly mobile loops β7-α7, β8-α8 are intrinsic to the structure upon only Phe binding.

Interestingly the dynamic properties of the distant Trp binding sites seem to be also affected by the binding of Phe. The residues on helix α1 and sheet β2 (residues 107-123) and helix α2a (residues 192-197) which constitute parts of the Trp binding pockets become less flexible. The residues on loop α2b-β3 (residues 237-242) which enclose the Trp binding sites become more flexible. This suggests that upon binding of Phe, part of the Trp binding site becomes more ordered favouring subsequent binding of Trp. In parallel to the behaviour of the N-terminal region in the Trp-only system, here in the presence of only Phe, the loop α2b-β3, which encloses the Trp binding sites, becomes more flexible to open up and facilitate entrance of the Trp molecules to the binding sites.

In summary, the simulations reveal that whereas ligand binding does not significantly alter the average conformations adopted by the *Mtu*DAH7PS tetramer, as also observed by the structural studies, it has a significant effect on the conformational fluctuations and the regional flexibilities of the protein, which may reflect the changes in the conformational energy landscape. For some regions of the enzyme, the relative energies of conformations around the average conformation become more similar, and the energy landscape around this average conformation can be considered to become somewhat flattened upon ligand binding. This alteration means that more conformations become energetically accessible, and as a result those corresponding regions of the enzyme become more flexible. It is proposed that this increased flexibility of the β2-α2 loop results in reduced affinity of the protein for substrate E4P, accounting for the inhibition of the enzyme by dual inhibitor binding. However, for other regions that become less flexible upon ligand binding, this suggests that the local minima on the conformational energy landscape become more pronounced, so that fewer conformations are energetically accessible. We observe this effect on the second inhibitor-binding site, when a single ligand is in place.

To date the majority of the known cases of allostery described involve gross conformational changes of the enzyme structure upon ligand binding. Computational and experimental techniques are now beginning to allow the

analysis of systems minimal average conformational changes are observed on ligand binding: *Mtu*DAH7PS with its complicated synergistic allostery appears to belong to this less studied group. The absence of distinct conformational states of the protein with altered functionality poses greater challenges for detecting the underlying molecular basis of the allosteric mechanism. A combination of computational and experimental techniques have been used to illustrate that the dynamic interplay between ligand binding sites is responsible for communication between allosteric sites and the synergistic allosteric regulation of *Mtu*DAH7PS. This study contributes to our increasing understanding of the allostery associated with limited conformational changes, and may help in elucidating a more general molecular understanding of the mechanisms for allosteric control.

6.3 SCA for type II DAH7PS family

The SCA method was conducted for type II DAH7PS family, in order to identify the co-evolved residues within this protein family that may be important for the observed synergistic allostery in *Mtu*DAH7PS. By applying this method to type II DAH7PS family, an assumption is made that allosteric regulation observed in *Mtu*DAH7PS is conserved between all members of this protein family. The SCA for type II DAH7PS family was conducted with MATLAB. All calculation algorithms, theories and program scripts used in the following analysis were developed by the Ranganathan lab (*109*). The SCA method was distributed by the original developers as a toolbox in the MATLAB station.

Step 1: obtaining and loading multiple sequence alignment and conditioning. All available amino acid sequences for the type II DAH7PS family were obtained from the KEGG encyclopaedia (*119-121*). A multiple sequence alignment (MSA) for all identified amino acid sequences of type II DAH7PS was carried out with CLUSTAL W (*122*). The produced MSA contains 262 sequences. This was imported into MATLAB for the SCA

method (*109*). To prevent trivial over-representation of gaps in the MSA and ensure that the calculations were only made at the largely non-gapped sequence positions, the sequence positions with gap frequencies greater than 20% were truncated from the full MSA. The truncated MSA contained 262 sequences with 433 positions in each amino acid sequence.

Step 2: sequence correlations. Ideally, the sequences in the MSA should represent a uniform sampling of the sequence space, in order to exclude the effect of biased sampling in the SCA. This could be examined from the structure of the correlations between sequences, through the computation of the matrix S of similarity between pairs of sequences, such that $S(s_1, s_2)$ gave the fraction of amino acids that were common between the sequences s_1 and s_2 in the truncated MSA.

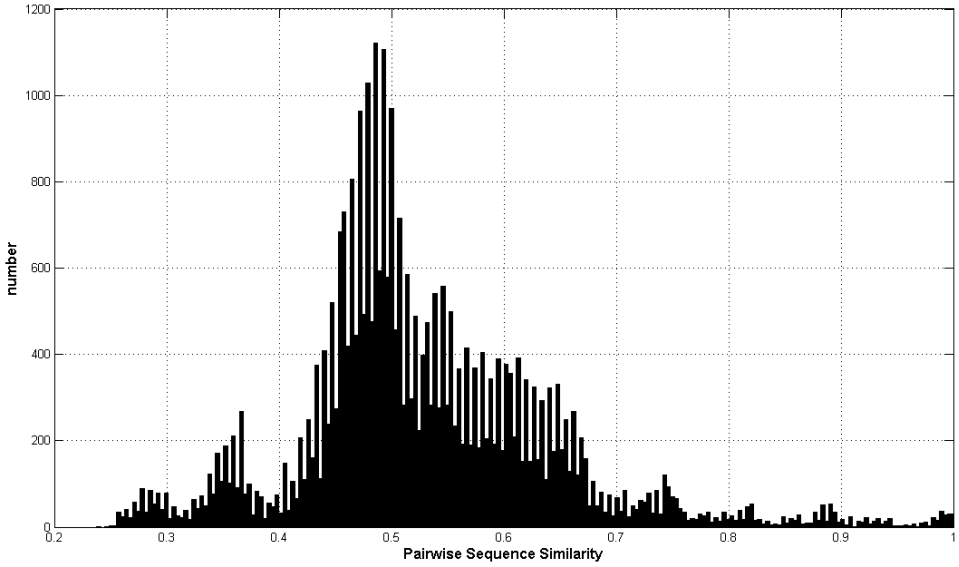


Figure 6. 6. Distribution of pairwise sequence similarities in the full set of MSA (262 amino acid sequences).

The distribution of sequence similarities computed in matrix S was visualised by a histogram (Figure 6. 6). The distribution showed the mean sequence similarity to be around 50%; however a number of sequences were shown to be highly similar (pairwise similarity above 90%). This high similarity was likely to cause sampling bias which will later affect the statistical analysis results. Therefore, those sequences with high similarities (above 80%) were removed from the full set MSA. This was carried out by inspection of the similarity matrix S, and the removal of groups of sequences with high

similarity score.

After the removal of a subgroup of high similarity sequences, the resultant MSA contains 69 amino acid sequences with 433 positions (the amino acid sequences used can be found in Appendix 5). To examine the sequence correlations in the refined MSA, the pairwise similarities between amino acid sequences in the refined MSA were calculated again by means of similarity matrix *S*. It is clear from this distribution of sequence similarities (Figure 6. 7) that all sequences with similarity above 80% were removed. The mean similarity is about 50%, and the similarity distribution ranges from 25% to 80% due to the small sequence divergence of this family. This distribution is not as ideal as the example system (the PDZ domain system) studied by the original developers of SCA (*116*), in which the MSA contains 240 sequences with mean similarity about 20% and similarity distribution from 5% to 50%. However, one of the most important requirements on the MSA used for the SCA method is the uniform distribution. After the removal of those highly similar sequences from the MSA of type II DAH7PS, the distribution of the amino acid sequences is considered reasonably uniform in the refined MSA and therefore suitable for the SCA method (Figure 6. 7). Another requirement on the MSA to be used in the SCA method is that the size of the sequence alignment should be large enough (~100 sequences). The refined MSA of type II DAH7PS family contains 69 sequences which is smaller than the required number. This means that there may be false positives identified in the final clusters of co-evolved residues due to statistical noise. Therefore, the interpretations of the final outcome from the SCA method should be considered carefully keeping in mind this limitation of MSA size.

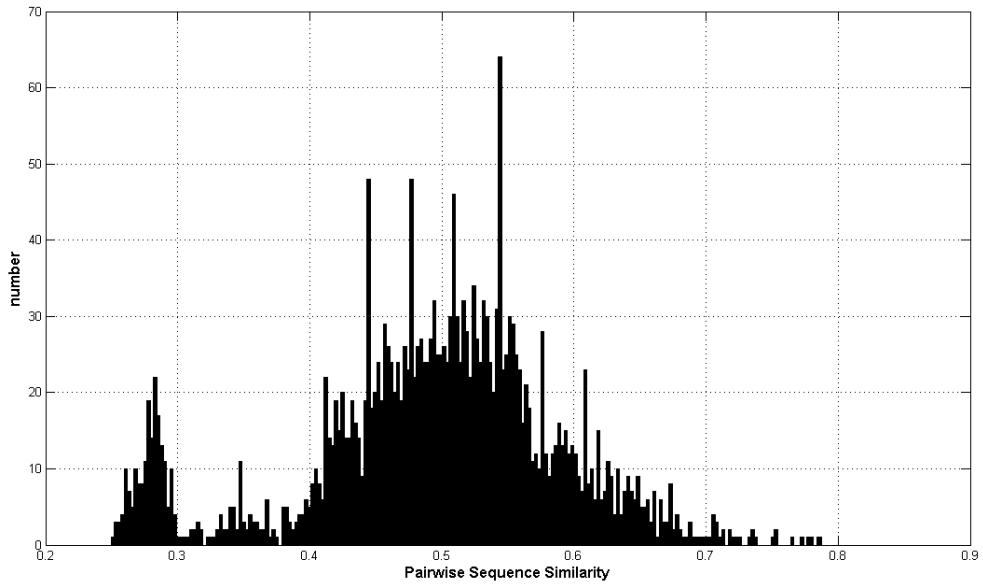


Figure 6. 7. Distribution of pairwise sequence similarities in the refined set of MSA (69 amino acid sequences).

Step 3: positional conservations. The degree of conservation of the different positions on the MSA was measured in SCA by an information theoretic quantity called the relative entropy, $D_i^{(a)}$. If $f_i^{(a)}$ represented the frequency of amino acid a at position i, and the probability of observing $f_i^{(a)}$ in an alignment of M sequences was represented by the binomial probability $\left(P_M\left[f_i^{(a)}\right]\right)$ with a given background probability of $q^{(a)}$, then the relative entropy $D_i^{(a)}$ was defined as:

$$D_i^{(a)} = -\frac{1}{M}\ln\left(P_M\left[f_i^{(a)}\right]\right)$$

This quantity indicated how unlikely the observed frequency of amino acid a at position i would be if it occurred randomly with probability $q^{(a)}$, and could capture the divergence of the observed frequencies of amino acids at each position from their background frequencies in the non-redundant database of proteins (*109*).

The positional conservation values (Figure 6. 8, top) of type II DAH7PS family show a mean value around 1.5 and the majority of the positions show conservations at a lower range from 0.5 to 3.5. Only a few positions show

high conservation values above 3.5. This pattern suggests that only the few functionally important residues are highly conserved in this protein family (*109, 123*).

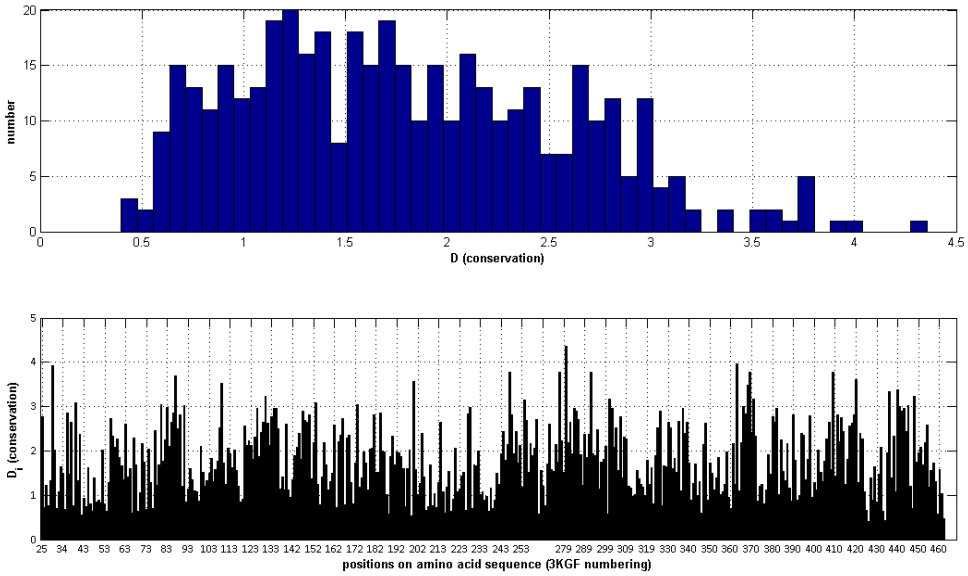


Figure 6. 8. Top: Distribution of positional conservation values of amino acids from the refined MSA of type II DAH7PS family. Bottom: Positional conservations plotted along the primary sequence of type II DAH7PS, the positions are numbered as in the MtuDAH7PS structure.

Plotting the positional conservation values along the primary amino acid sequence shows a dispersed pattern (Figure 6. 8, *bottom*). However, mapping the conservation values on to three dimensional crystal structure of *Mtu*DAH7PS (Figure 6. 9B) shows that most of the positions with high conservation values are located at the centre of the barrel structure. They constitute part of the active site and are important for maintaining the catalytic function of the enzyme. Some of them are located on the β-strands of the core barrel structure, which are important for maintaining the basic fold of this protein family. Several other positions with high conservation values were found at the dimer interface (Figure 6. 9A), e.g. Pro56, Arg181 and Arg184, and tetramer interfaces, e.g. Phe227.

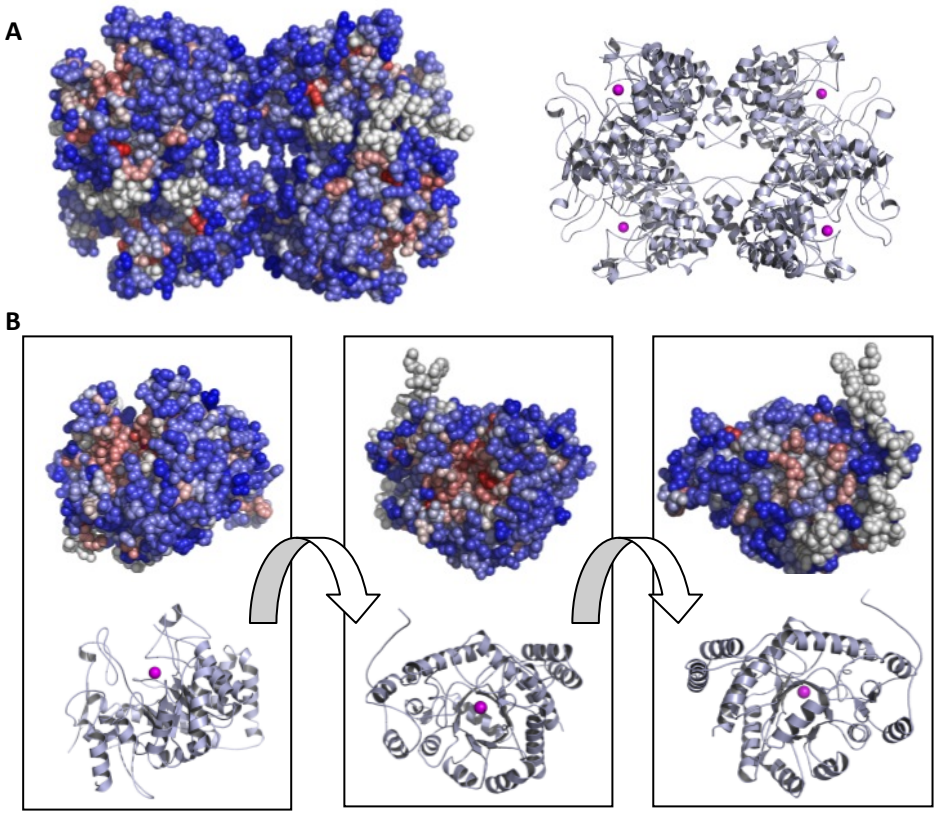


Figure 6. 9. Positional conservation values mapped on crystal structure of *Mtu*DAH7PS. The conservation values are represented using a colour scale, blue indicates low conservation and red indicates high conservation. A. Positional conservation values mapped on the tetramer of *Mtu*DAH7PS, the tetramer structure in the same orientation with cartoon representation is shown on the right to indicate positions of active site and interfaces. B. Positional conservation values mapped on the monomer of *Mtu*DAH7PS shown in three different orientations (highlighted in three boxes), and the monomer structure in each orientation is displayed by both sphere form and ribbon form, to indicate the position of the active site.

Step 4: positional correlations. The co-evolved residues within a protein family were identified in the SCA by calculating the correlation of positional conservations between each pair of positions on the truncated refined set of MSA. The correlation values were represented in a correlation matrix ($C_{ij}^{(ab)}$) which is defined as:

$$C_{ij}^{(ab)} = f_{ij}^{(ab)} - f_i^{(a)} f_j^{(b)}$$

Where $f_{ij}^{(ab)}$ represented the joint frequency of having amino acid a at position i and amino acid b at position j. Here, the correlation matrix for type II DAH7PS family was calculated with the "binary approximation" applied. In the "binary approximation", only the most frequent amino acid a_i at position i was considered. The alignment was then represented by a

binary array $x_{i,s}$ where $x_{i,s} = 1$ if sequence s contained the most frequent amino acid at position i, and 0 otherwise. After application of the binary approximation, the correlation matrix $C_{ij}^{(ab)}$ was reduced to:

$$C_{ij} = f_{ij}^{(a_i a_j)} - f_i^{(a_i)} f_j^{(a_j)}$$

This reduction was proved to be useful by the original developers of the SCA method (*109*), since the positional conservation in the full alignment was well-approximated by the positional conservation in the binary alignment.

The full correlation matrix for type II DAH7PS family was plotted using a colour scale (Figure 6. 10, *top*), where blue indicated low correlations and red indicated high correlations between pairs of positions on the MSA, and the x and y axes represented the positions on the MSA. Most of the pairs of positions showed low correlations (blue) to each other. However, a few non-adjacent pairs of positions (off-diagonal) with medium or high correlations can still be observed. This was clear from the distribution of correlation values (Figure 6. 10, *bottom*), which was skewed with a long tail into the high correlation range. Due to a large number of positions in the MSA, it is difficult to identify exactly which positions were showing high correlations by inspection of the correlation matrix. Therefore, a clustering analysis was carried out to reveal more clearly the statistically coupled positions.

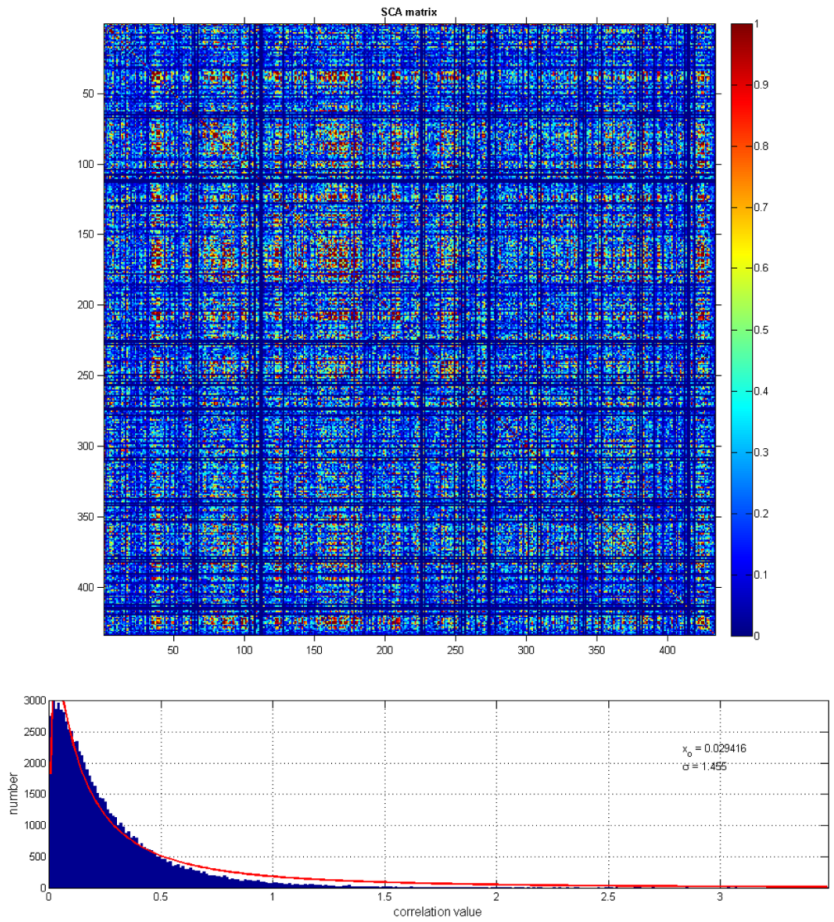


Figure 6. 10. Top: Full SCA matrix for type II DAH7PS family, correlation values are shown using a colour scale, blue indicates low correlation and red indicates high correlation. Bottom: Distribution of the positional correlation values.

Step 5: clustering analysis. In this clustering analysis, positions in the correlation matrix were grouped into clusters according to their correlation values. The clusters were defined so that the most closely related positions are placed next to each other. The correlation values were plotted onto a new matrix in which the positions on the x and y axes were "scrambled" according to the clustering order. The cluster regions of high correlation values were identified from the new, sorted correlation matrix, based on visual inspection of the sorted correlated matrix and the clustering tree, and further iterations of clustering analysis were carried out until the positions in the identified clusters converged.

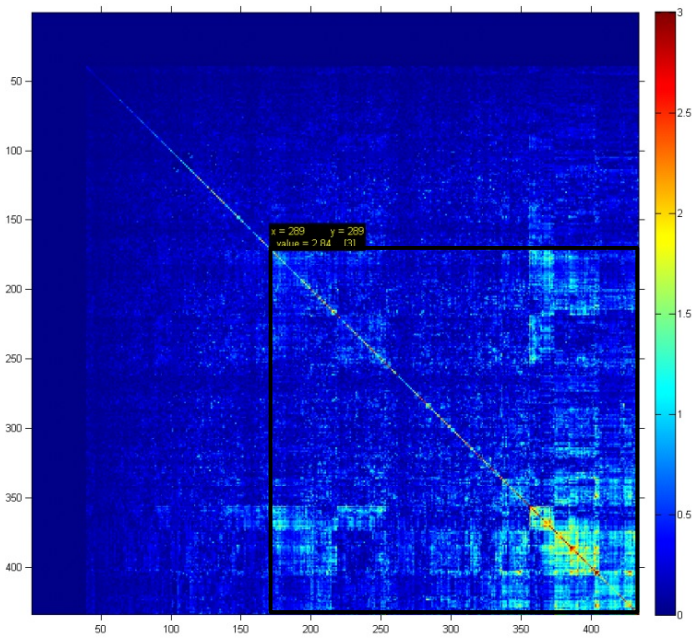


Figure 6. 11. Sorted correlation matrix after the first iteration of clustering analysis. The correlation values between pairs of positions are shown using a colour scale, with blue indicating low correlations and red indicating high correlations. The x and y axes represent positions on the MSA, but the order of these positions are sorted according to the clustering. The region identified for further iterations of clustering analysis was highlighted in black box

- *Iteration 1*

After the first iteration of clustering analysis on the correlation matrix, the correlations were sorted into clusters and the matrix was reordered accordingly (Figure 6. 11).

From this, a region of high correlation values was identified (Figure 6. 11). This region was selected to go through a second iteration of the clustering analysis, and it included the following positions (with *Mtu*DAH7PS numbering):

289+26+202+386+45+350+90+446+102+388+401+49+220+64+237+187+428+98+364+236+352+37
4+241+274+279+105++160+191+30+251+151+169+244+122+225+361+172+213+385+273+276+45
2+457+286+440+438+394+29+132+307+389+437+393+387+68+357+152+93+183+334+36+37+95+
316+174+342+418+353+408+195+324+327+329+449+222+283+344+405+115+224+333+403+33+7
9+303+34+156+104+238+295+55+203+159+392+373+377+70+170+317+320+211+427+148+254+2
91+368+331+74+272+168+83+308+246+281+96+137+217+138+258+332+416+52+314+400+319+3
23+81+124+82+413+109+455++154+180+299+318+345+142+340+176+321+38+417+111+209+379
+214+395+41+293+48+205+343+431+287+365+72+354+165+397+410+458+201+338+58+432+179
+146+164+63+260+177+107+194+259+99+192+193+143+158+378+384+243+25+28+298+40+245+
294+297+271+103+131+65+94+175+101+412+100+262+278+57+108+84+144+162+145+112+451+
147+185+157+60+186+61+62+189+204+110+218+114+435+123+232+182+197+230+231+391+198
+200+228+454+59+423+406+71+275+372+450+67+184+128+381+414+422+348+121+221+249+30
9+120+453+181+190+227+173+358+402+363+398

A subset of the full SCA correlation matrix was made to include only positions identified above. This smaller correlation matrix was subjected to the second round of clustering analysis.

- *Iteration 2*

After the second iteration of clustering analysis on the subset of the correlation matrix, the correlations were sorted into clusters again and the matrix was reordered accordingly (Figure 6. 12). From this sorted correlation matrix, two regions of high correlation values were observed (Figure 6. 12). These two regions contained the following positions (*Mtu*DAH7PS numbering):

67+184+128+381+348+414+422+121+58+432+179+99+192+63+260+177+107+158+384+243+160+191+452+457+30+251+273+276+231+391+52+81+124+59+423+406+372+71+275+450+72+354+165+397+410+458+194+259+193+221+249+309+289+26+202+297+324+90+446+105+271+386+102+388+401+45+350+49+220+428+374+151+64+237+187+55+98+364+236+352+241+274+279+103+131+146+164+245+294+25+28+298+40+65+94+175+101+412+100+262+278+143+378+358+402+398+120+453+181+190+227+173+363+57+108+84+144+145+162+112+451+147+157+185+230+189+204+198+60+186+61+62+110+218+435+114+123+232+182+197+200+228+454

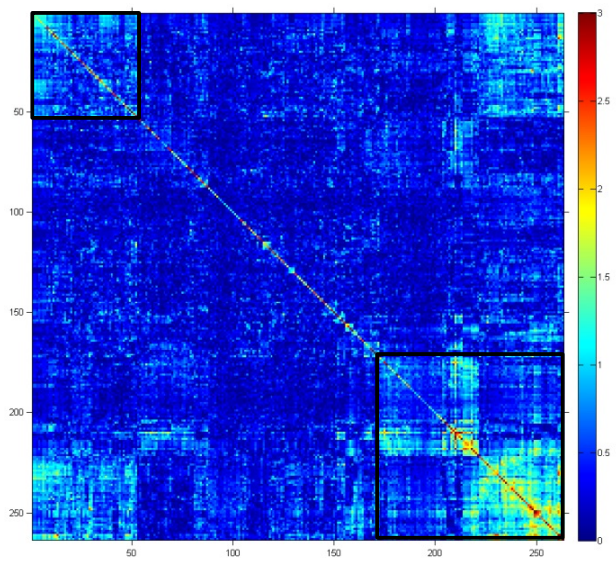


Figure 6. 12. Sorted correlation matrix after the second iteration of clustering analysis. The correlation values between pairs of positions are shown using a colour scale, with blue indicating low correlations and red indicating high correlations. The x and y axes represent positions on the MSA, but the order of these positions are sorted according to the clustering. The regions identified for further iterations of clustering analysis were highlighted in black boxes.

The same procedure was applied as before, that a subset of the correlation matrix was generated with only positions identified from the above two

regions in the second iteration. This subset of the correlation matrix was subjected to the third round of clustering analysis.

- *Iteration 3*

After the third iteration, three regions of high correlation values were observed (Figure 6. 13), these regions contained the following residues (*Mtu*DAH7PS numbering):

94+175+101+412+241+274+279+103+131+164+294+245+298+146+40+25+28+65+100+262+173+143+453+378+120+181+402+398+363+57+108+157+185+189+204+198+230+60+182+26+202+162+112+147+110+446+324++388+186+84+451+144+145+271+102+386+401+278+358+227+61+62+105+218+435+114+123+232+197+200+228+454+128+381+422+348+414+121+251+64+428+374+151+350+49+220+45+55+98+364+236+352

A subset of the correlation matrix containing only the above positions was generated and subjected to the fourth iteration of clustering analysis.

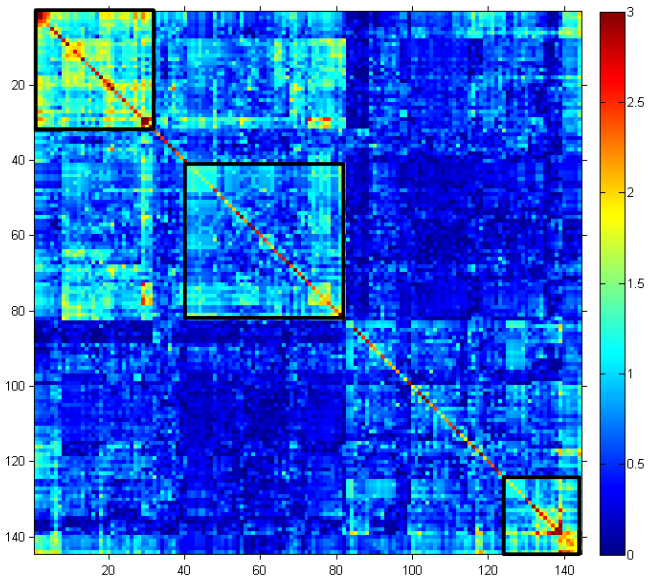


Figure 6. 13. Sorted correlation matrix after the third iteration of clustering analysis. The correlation values between pairs of positions are shown using a colour scale, with blue indicating low correlations and red indicating high correlations. The x and y axes represent positions on the MSA, but the order of these positions are sorted according to the clustering. The regions identified for further iterations of clustering analysis were highlighted in black boxes.

- *Iteration 4*

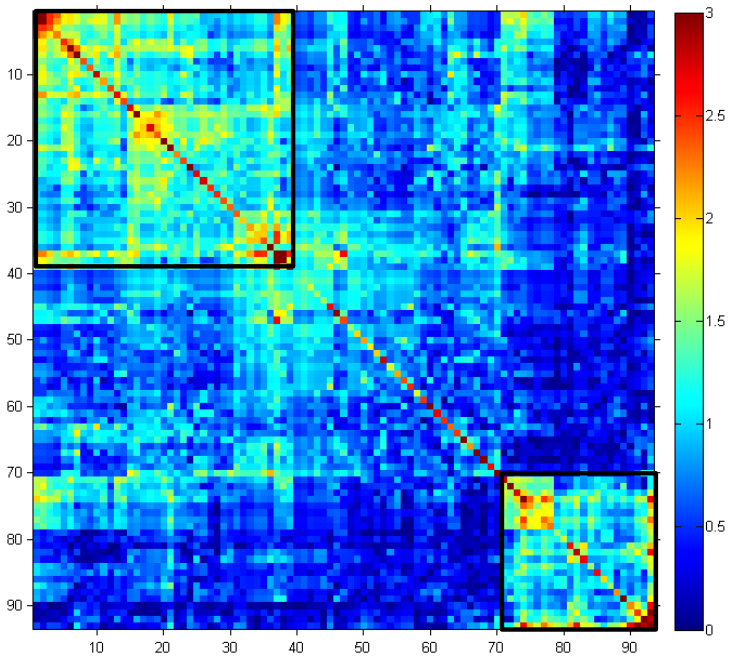


Figure 6. 14. Sorted correlation matrix after the fourth iteration of clustering analysis. The correlation values between pairs of positions are shown using a colour scale, with blue indicating low correlations and red indicating high correlations. The x and y axes represent positions on the MSA, but the order of these positions are sorted according to the clustering. The regions identified for further iterations of clustering analysis were highlighted in black boxes.

The fourth iteration of clustering analysis identified two regions of high correlation values (Figure 6. 14). These two regions contained the following positions (*Mtu*DAH7PS numbering):

94+175+101+412+100+262+173+143+453+378+120+181+402+398+103+131+164+294+245+298+65+146+40+358+25+28+271+102+197+200+218+435+114+123+232+228+363+57+108+241+279+274+352+55+236+98+364+128+381+151+428+374+422+121+348+414+251+64+350+49+220+45

After the fourth iteration, the position numbers identified from high correlation clusters started to converge, therefore, a final iteration of clustering analysis was carried out to finalise the definition of clusters for the type II DAH7PS family. A subset of the correlation matrix was generated containing only the positions identified above, and this matrix was subjected to the final round of clustering analysis.

- *Iteration 5 (final iteration)*

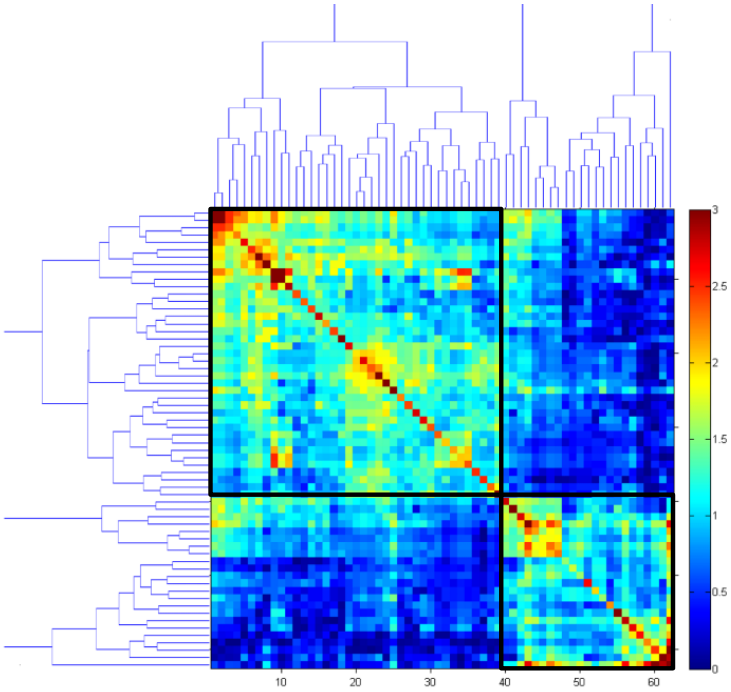


Figure 6. 15. Sorted correlation matrix after the final iteration of clustering analysis. The correlation values between pairs of positions are shown using a colour scale, with blue indicating low correlations and red indicating high correlations. The x and y axes represent positions on the MSA, but the order of these positions are sorted according to the clustering. The clustering tree was displayed on the top and left of the correlation matrix. The final clusters identified for type II DAH7PS family were highlighted in black boxes.

The final iteration of clustering analysis identified two clusters of positions with high correlation values (Figure 6. 15). These two clusters were consistent with the two clusters found after iteration 4, indicated that the clustering analysis had converged. A total of 62 positions were identified as co-evolved residues within the type II DAH7PS family. The two clusters contained the following positions (*Mtu*DAH7PS numbering):

Cluster 1:

94+175+101+412+402+100+262+173+363+57+108+143+453+398+378+120+181+358+103+164+294+245+131+298+65+146+40+28+25+197+218+435+114+123+232+271+102+200+228

Cluster 2:

241+279+274+352+55+236+98+364+128+381+422+121+348+414+251+64+428+374+151+350+49+220+45

These two clusters of co-evolved residues were mapped onto the crystal structure of *Mtu*DAH7PS, to examine the locations of these residues on the tertiary structure of the enzyme, and to investigate the possible physical meaning of these clusters.

The first cluster contains 39 co-evolved residues with high correlation values (Figure 6. 16). Most of these co-evolved residues are found on the α-helices which are on the exterior of the barrel structure. This is consistent with the results about the conservation values discussed above that the centre of the barrel contains most of the highly conserved positions in the type II DAH7PS family (Figure 6. 9), since the highly conserved residues usually have very low correlation values and do not show up in the correlation analysis. The first cluster of co-evolved residues can be found at the Phe binding sites (e.g. residues 94 and 175), at the Trp binding sites (e.g. residues 123, 114 and 197), and on the loops in the active site (e.g. residues 143, 146, and 378). The other co-evolved residues in cluster 1 are found to mediate physical connectivity through van der Waals interactions between these three distant sites. The communication signal between the Phe and Trp binding sites within the same monomer could propagate along those co-evolved residues on helices α1 and α2 (Figure 6. 16c). The ligand binding signals could potentially propagate across different monomers through those co-evolved residues at the dimer and tetramer interfaces (e.g. residues 228, 181).

Therefore, the first cluster of co-evolved residues within the type II DAH7PS family may represent the communication pathway between the allosteric ligand binding sites and the active site, throughout the quaternary (tetrameric) structure. Ligand binding at the distant allosteric sites at the dimer and tetramer interfaces may first disrupt the movement of those residues in their binding sites, and this disruption signal can then propagate along the pathway formed by these co-evolved residues in the first cluster, and eventually reach the active sites in order to achieve inhibition of the enzyme.

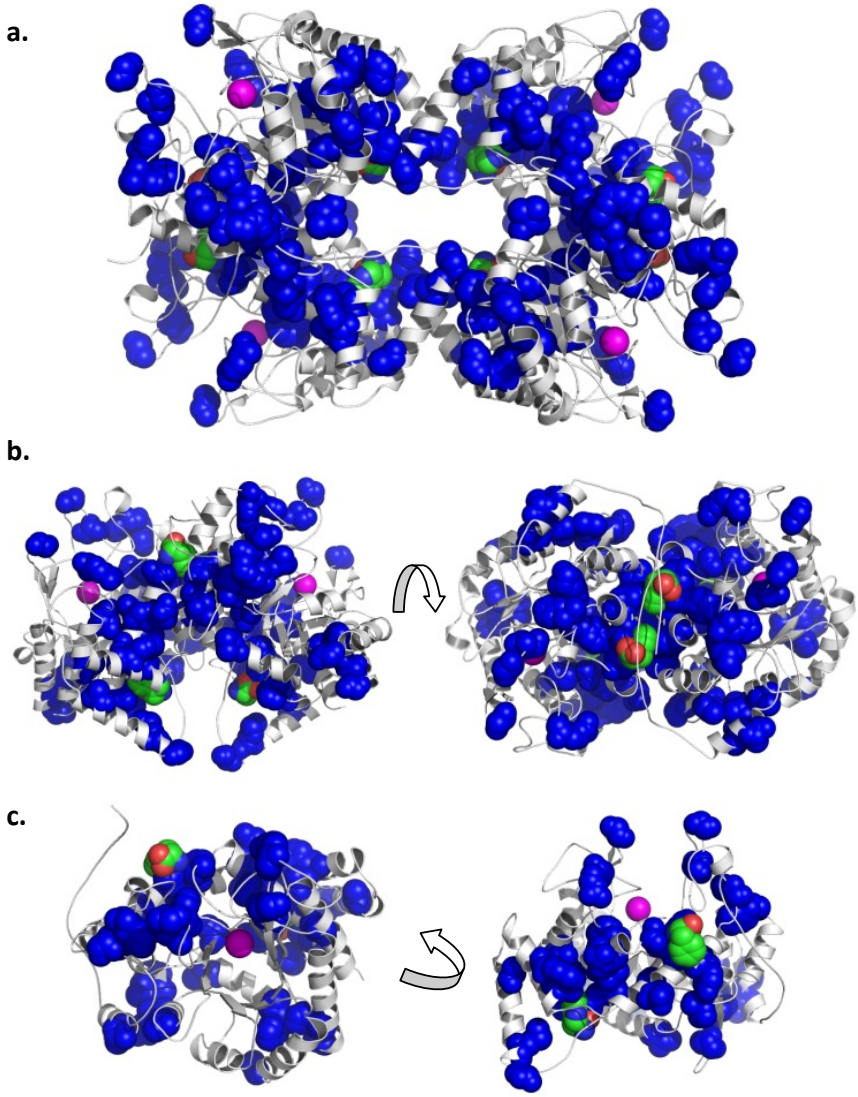


Figure 6. 16. Co-evolved residue in the first cluster identified from the SCA for type II DAH7PS family, mapped onto the ligand-bound (Trp+Phe) crystal structure (PDB 3KGF) of *Mtu*DAH7PS, shown with white cartoon representation. The tetramer, dimer and monomer structures are shown in a, b, and c respectively. The co-evolved residues are displayed as blue spheres; the bound Trp and Phe ligands are displayed with green carbon atoms. The active site metal ions, Mn^{2+}, are displayed as magenta spheres to indicate the position of the active site.

The second cluster contains 23 co-evolved residues (Figure 6. 17). Similar to cluster 1, most of these co-evolved residues are found on the outside of the core barrel structure. These co-evolved residues are found between helices and sheets within the same monomer (e.g. residues 151, 251, 64 and 274) and between different monomers at the dimer and tetramer interfaces (e.g. residues 220, 236). Since these co-evolved residues are not located at the ligand binding sites, they may represent the communications between secondary structural elements in the enzyme in the absence of the inhibitors.

These residues may have co-evolved due to their importance in maintaining the basic barrel fold of the enzyme, and/or maintaining the functional quaternary structures within the type II DAH7PS family.

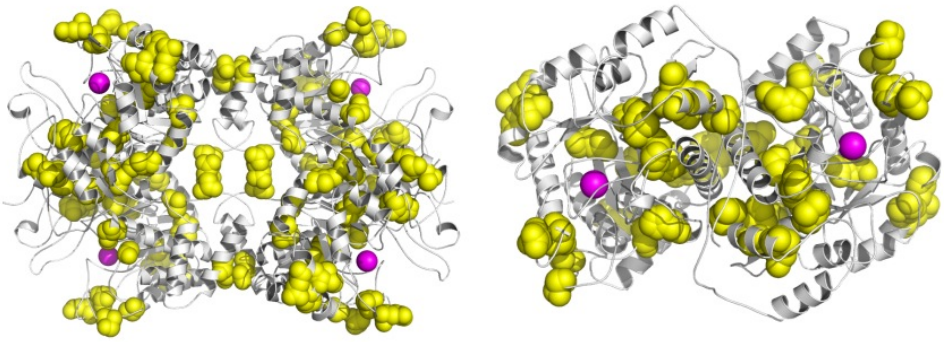


Figure 6. 17. Co-evolved residues in the second cluster identified from the SCA for type II DAH7PS family, mapped onto the crystal structure (PDB 3KGF) of *Mtu*DAH7PS, shown with white cartoon representation. The tetramer, and dimer structures are shown on the left and right respectively. The co-evolved residues are displayed as yellow spheres. The active site metal ions, Mn^{2+}, are displayed as magenta spheres to indicate the position of the active site.

In summary, SCA was carried out for type II DAH7PS family, in order to identify important co-evolved residues and possible communication pathways between distant allosteric inhibitor binding sites and the active sites. The analysis identified 62 co-evolved residues in two clusters. The first cluster of co-evolved residues can be found to mediate propagation pathway between the allosteric sites and active sites throughout the quaternary structure. This cluster is considered to represent the communication pathway between the three distant sites, and is of potential interest in deciphering the allosteric communication mechanism. The second cluster of co-evolved residues may be important for structural reasons.

6.4 SCA·MD analysis for *Mtu*DAH7PS

The SCA method identified two clusters of residues (total of 62 residues) that co-evolved within the type II DAH7PS family, of which *Mtu*DAH7PS is a member. The SCA proposed a large number of potentially interesting targets for further investigation by experiments. In order to narrow down the range

of potential targets, MD simulations for wild type *Mtu*DAH7PS at the ligand-free state, and in combination with different allosteric inhibitors were carried out. The movements of the residues were analysed and combined with the results in SCA to filter out those co-evolved residues that were important for structural reasons, and to propose reasonable numbers of candidates for further experimental investigations.

6.4.1 Residue movement correlations from MD simulations

The atomic fluctuations of each amino acid of the tetramer of *Mtu*DAH7PS during the MD simulations are represented by the average displacement vectors (Δr). The displacement vectors were measured using the first frame of the equilibrated part of the trajectory (2.3 ns to 20 ns) as the reference, and all non-hydrogen atoms of each residue (both backbone and side chain), were included for the calculation. Then the correlations of movement between a pair of residues i and j is represented by the correlation coefficient ($R_{i,j}$) of the covariance matrix ($COV_{(\Delta r_i, \Delta r_j)}$) between Δr_i and Δr_j, where the covariance matrix is defined as:

$$COV_{(\Delta r_i, \Delta r_j)} = C_{(i,j)} = E\left[\left(\Delta r_i - \mu_{r_i}\right)\left(\Delta r_j - \mu_{r_j}\right)\right]$$

in which E is the mathematical expectation and $\mu_{r_i} = E_{r_i}$. Then the correlation coefficient ($R_{i,j}$) (i.e. the correlation of motion between pair of residues i and j) is calculated by:

$$R_{i,j} = \frac{C_{(i,j)}}{\sqrt{C_{(i,i)} C_{(j,j)}}}$$

The correlation coefficient values ($R_{i,j}$) range from -1 to 1. Positive correlation values indicate that the pair of residues moves in the same direction, whereas the negative correlation values indicate the residues move in the opposite direction. A completely correlated ($R_{i,j} = 1$) or anti-correlated ($R_{i,j} = -1$) motion means that the motions have the same phase and period.

The matrix of the correlation coefficients (R) was determined based on MD trajectories from each of the four systems of *Mtu*DAH7PS (ligand-free,

Trp+Phe, Trp-only, and Phe-only bound), in order to detect any difference in correlations of motion upon ligand binding.

Ligand-free *Mtu*DAH7PS. The correlation coefficient matrix (R) for the ligand-free *Mtu*DAH7PS was plotted using a colour scale (Figure 6. 18) where blue indicates complete anti-correlations and red indicates complete correlations between the corresponding pair of residues.

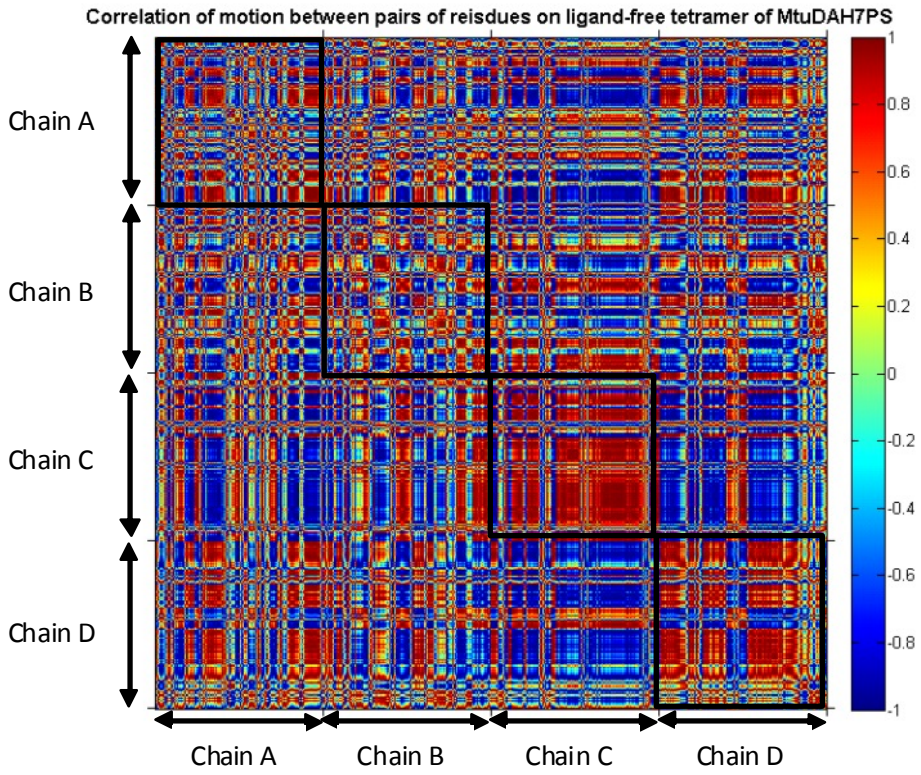


Figure 6. 18. Correlation matrix (R) of residue motions during MD simulations of ligand-free *Mtu*DAH7PS tetramer. The correlation values were plotted using a colour scale, with red indicating complete correlation in motion, and blue indicating complete anti-correlation in motion.

From the tetramer correlation matrix, two sets of information can be obtained, including the correlations of motion between residues on the same monomer (highlighted in black boxes in Figure 6. 18), and correlations of motion between different monomers. The same-monomer-correlations can provide an indication of the relative movement between specific loops and helices within the same monomer, whereas the different-monomer-correlations can provide an indication of which monomer is moving in correlation (or

anti-correlation) with the other monomers, thus provide a picture of overall domain movements on the tetramer of *Mtu*DAH7PS.

The four monomers in the ligand-free tetramer complex show different same-monomer-correlation patterns, indicating that the relative movements of the secondary structure elements are different in each monomer. Chains A and B show similar regions of correlation and anti-correlation. Chains C and D show larger regions of correlated motion, especially for chain C, in which almost all the regions on the monomer are moving in correlation, with the exception of the N-terminus, and several loop structures.

The different-monomer-correlations show that chain D was mostly moving in correlation with chain A and in anti-correlation with chain C.

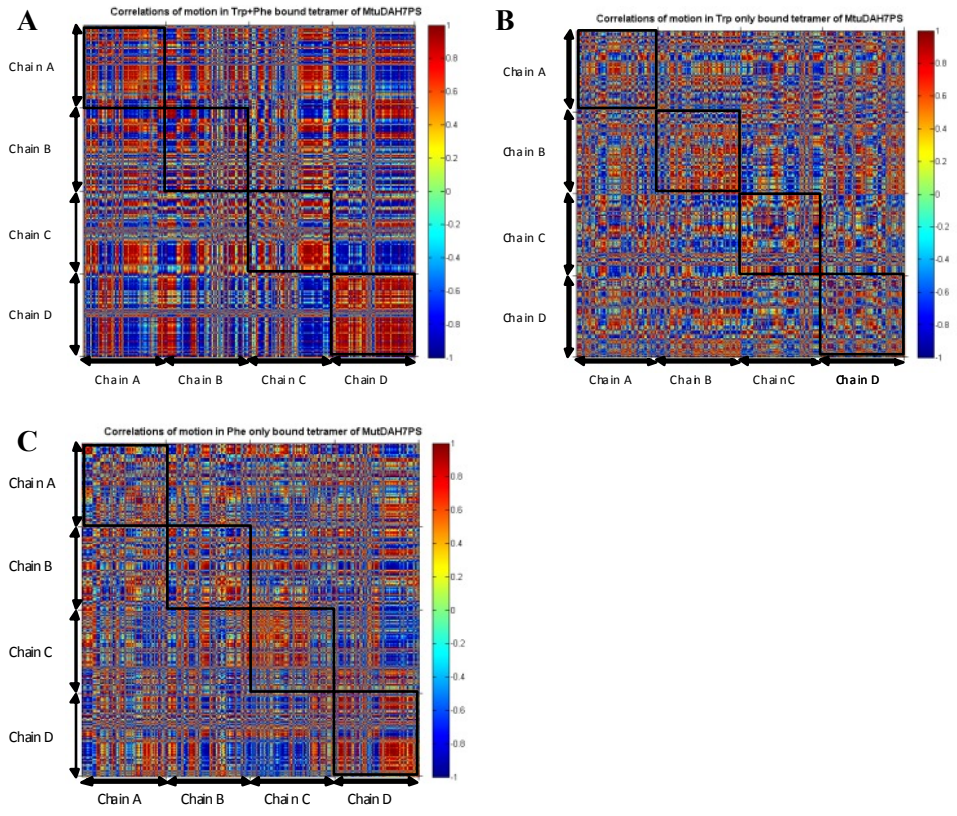


Figure 6. 19. Correlation matrix (R) of residue motions during MD simulations of A)Trp+Phe bound, B) Trp-only bound and C) Phe-only bound *Mtu*DAH7PS tetramer. The correlation values were plotted using a colour scale, with red indicating complete correlation in motion, and blue indicating complete anti-correlation in motion.

Ligand-bound *Mtu*DAH7PS. The same correlation matrix was calculated for the three ligand-bound systems of *Mtu*DAH7PS tetramer (Trp+Phe, Trp-only and Phe-only bound), and plotted using the same colour scale (Figure 6. 19).

The three ligand-bound systems of *Mtu*DAH7PS tetramers show different correlations of motion during the MD simulations (Figure 6. 19). When both Trp and Phe are bound, the relative domain movement is altered compared to the ligand-free system, but the regions of correlation and anti-correlation are still clearly defined from the matrix. On the other hand, in the two single ligand-bound systems (Trp-only and Phe-only), the patterns of correlation and anti-correlation are completely disrupted, and the regions are not as clearly defined as in the ligand-free system. These differences suggest that binding of single or double ligands can alter the relative movements between secondary structure elements within the same monomer, as well as the relative movements between different monomers on the tetramer of *Mtu*DAH7PS. These differences in correlations may be induced by ligand binding to the *Mtu*DAH7PS tetramer, and may contribute to the mechanism of the propagation of ligand binding signals throughout the tetramer structure.

6.4.2 SCA·MD hybrid analysis for Mtu*DAH7PS*

As mentioned previously, the SCA was used to identify groups of residues that have evolved in correlation with each other within the type II DAH7PS family. The identified co-evolved residues were important for either functional or structural reasons. Here, MD simulations were conducted, and residues on the *Mtu*DAH7PS tetramer that have moved in correlation (or anti-correlation) were identified by computing the MD correlation matrix.

Then a hybrid matrix (SCA·MD) can be computed by multiplying the individual elements of the SCA matrix with the corresponding elements of the MD matrix. The resultant hybrid matrix contains information from both the MD simulation and the SCA. In this calculation, the elements from the

MD matrix can be considered to act as a "weighting" term for the SCA matrix, so that those co-evolved residues from the SCA that are only important for maintaining basic fold of the enzyme will be down-weighted. As a result, the elements with high correlation values from the hybrid matrix belong to a subset of the co-evolved residues, which contains the ones that are important for the dynamics fluctuations within this enzyme.

From analysis of the crystal structures of *Mtu*DAH7PS in complex with different ligands (Chapter 4), it can be observed that there is no major conformational change upon ligand binding. Therefore, the synergistic allosteric inhibition of *Mtu*DAH7PS by binding of ligands (Trp and Phe) was predicted to be contributed mainly by altering the dynamics within the structure upon ligand binding. Dynamic pathways between the allosteric ligand binding sites and the active sites must exist for the ligand binding signal to be transmitted to achieve inhibition. Therefore, residues with high correlation values in the hybrid matrix (SCA·MD) were identified in order to find the dynamic pathways.

SCA·MD analysis for ligand-free *Mtu*DAH7PS. The MD correlation matrix contains both positive and negative values, indicating correlation and anti-correlation in motion respectively. In this hybrid analysis however, the phase of the correlation is not important, since what is of interest is the magnitude of the correlation values between pairs of residues. Anti-correlation can also be considered as a special case of correlation, i.e. pairs of residues with large anti-correlation values are the same as pairs of residues with large correlation values but in the opposite direction. The SCA matrix does not contain phase information either. Therefore, the absolute values of the elements in the MD matrix were obtained and this absolute MD matrix with all positive values was used to compute the hybrid matrix. The MD matrix has dimension 1848 by 1848, because each monomer on the structure was observed to behave differently during the dynamics simulations, and in order to discover the dynamic pathway within the tetramer complex, all of the four monomers must therefore be considered. The SCA matrix however only has dimension 462 by 462, since it is solely based on the

amino acid sequence of the enzyme, and this enzyme is a homo-tetramer, i.e. all of the four monomers share the same amino acid sequence. To make the MD matrix and the SCA matrix consistent in dimension, a "tetrameric" form of the SCA matrix (1848 by 1848) was generated. The tetrameric SCA matrix was then normalised so that it has a maximum value of 1, to make sure that the SCA matrix is not over-emphasised in the final hybrid matrix.

The SCA·MD hybrid correlation matrix for ligand-free *Mtu*DAH7PS was plotted using a colour scale (Figure 6. 20), where blue indicated low correlations and red indicated high correlations between pairs of residues on the tetramer, and the x and y axes represented the positions on the tetramer structure of *Mtu*DAH7PS. Most of the pairs of residues show low correlations (blue) to each other. However, a few non-adjacent pairs of positions (off-diagonal) with medium or high correlations can still be observed.

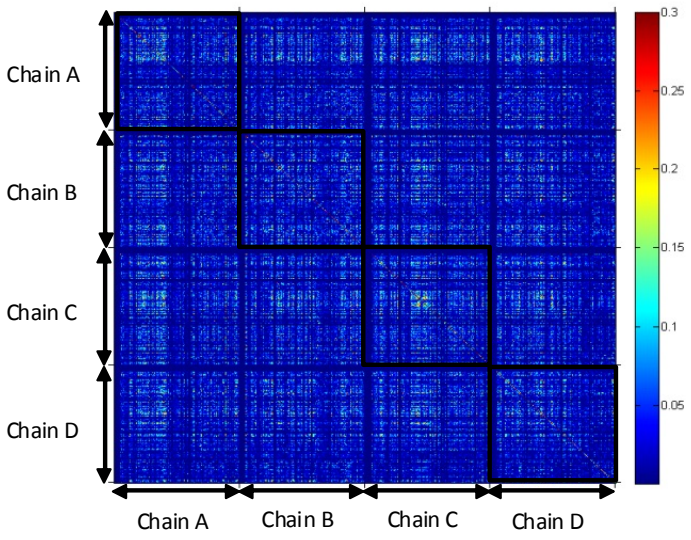


Figure 6. 20. SCA·MD matrix for ligand-free *Mtu*DAH7PS tetramer plotted using a colour scale. Blue indicates low correlation and red indicates high correlation.

As with the SCA method, a clustering analysis was carried out to analyse the hybrid matrix, in order to reveal more clearly the coupled amino acids (details of clustering iterations are not shown). The final iteration of clustering analysis identified two clusters of positions with high correlation

values (Figure 6. 21). A total of 169 positions were identified. The two clusters contained the following positions:

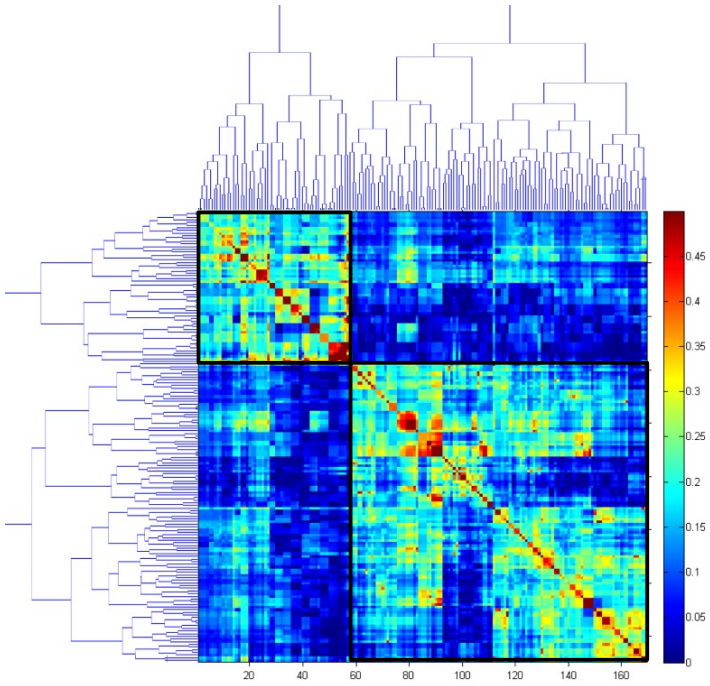


Figure 6. 21. Sorted SCA·MD matrix after the final iteration of clustering analysis. The correlation values between pairs of positions are plotted using a colour scale, with blue indicating low correlations and red indicating high correlations. The x and y axes represent positions on the *Mtu*DAH7PS tetramer, but the order of these positions are sorted according to the clustering. The clustering tree was displayed on the top and left of the correlation matrix. The final clusters identified for *Mtu*DAH7PS were highlighted in black boxes.

Cluster 1:

90B+90C+90D+90A+446A+446B+101A+101D+65C+175A+101C+94D+1
75D+100A+278C+278D+412A+412D+412B+65A+175B+94B+262A+262
C+262D+262B+412C+271C+271D+294A+294C+294D+245B+245C+245D
+297A+297C+297D+297B+40B+40C+40D+131A+131D+131B+131C+103
A+103D+103C+298A+298C+298D+298B+25A+25C+25D+25B

Cluster 2:

157C+185D+185A+57A+57C+358D+112B+57D+108B+162C+453A+453B
+120B+414C+363A+84B+84C+84D+62C+61A+61C+60A+60C+186B+186
D+123B+190A+190C+190D+454A+454B+200A+200D+200B+200C+218
A+114D+185B+110B+198D+197B+198A+204B+204D+197A+182C+60D
+61D+57B+454D+192B+192D+99B+99D+157A+451A+451B+144A+162
A+147A+182D+182B+232A+232B+114B+218D+435A+197C+197D+123
A+123C+198C+108C+189A+189B+189D+204C+112A+112C+181A+190B
+414A+453C+120A+120C+227A+227B+454C+228A+228D+228B+228C+

451C+147C+147D+145A+145B+145D+144B+162B+358B+358C+402C+3
63C+363D+145C+144D+108A+108D+162D+157B+157D

The residues in each cluster were from different chains, to help clarify the residues from each chain that were identified, the residues were sorted according to the chains in each cluster. The SCA·MD analysis of the ligand-free *Mtu*DAH7PS identified the following two clusters of residues with high correlation values:

Cluster 1:

Chain A →
90+25+262+297+131+298+412+175+446+65+101+100+294+103
Chain B →
90+25+262+297+131+298+412+175+446+94+245+40
Chain C →
90+25+262+297+131+298+412+65+101+294+103+245+40+278+271
Chain D →
90+25+262+297+131+298+412+175+101+294+103+94+245+40+278+271

Cluster 2:

Chain A →
57+190+200+157+454+197+162+228+145+108+185+453+363+61+60+218
+198+451+144+147+
232+435+123+189+112+181+414+120+227
Chain B →
57+190+200+157+454+197+162+228+145+108+185+453+144+232+123+1
89+112+120+227+84+186+110+204+192+99+451+182+114+358
Chain C →
57+190+200+157+454+197+162+228+145+108+453+363+61+60+198+451
+147+123+112+414+
120+84+204+182+358+62+402
Chain D→
57+190+200+157+454+197+162+228+145+108+185+363+61+60+218+198
+144+147+189+84+186+192+99+182+114+358+204

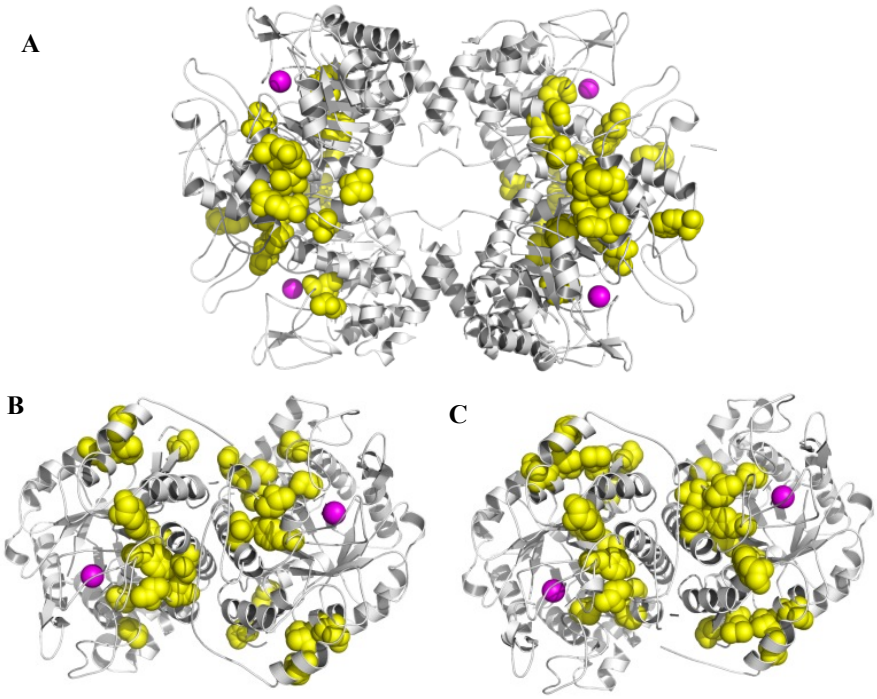


Figure 6. 22. Cluster 1 from SCA·MD analysis of ligand-free *Mtu*DAH7PS. The enzyme is displayed in white cartoon representation; the metal ions Mn^{2+} are shown as magenta spheres to indicate positions of the active sites. The residues found in cluster 1 of ligand-free MtuDAH7PS are displayed as yellow spheres on tetramer (A), chains A and B (B), and chains C and D (C) of MtuDAH7PS.

These two clusters of highly correlated residues were mapped onto the crystal structure of ligand-free *Mtu*DAH7PS, to examine the locations of these residues on the tertiary (and quaternary) structure of the enzyme, and to investigate the possible physical meaning of these clusters.

The first cluster contains 57 residues on the tetramer of ligand-free *Mtu*DAH7PS (Figure 6. 22). Most of these correlated residues were found between helices within the two dimers, and no connectivity was found across the tetramer interface. This cluster of residues may be responsible for controlling the relative motion of structural elements within the same monomer through van der Waals interactions.

The second cluster contains 112 correlated residues on the ligand-free tetramer of *Mtu*DAH7PS (Figure 6. 23). Most of these residues are found across the tetramer interface (e.g. residues 227, 228 and 120), within the Trp binding sites (e.g. residues 123 and 197), across the dimer interface (e.g. residues 181 and 184), and some are found on the active site loops (e.g.

residues 147, 145, 157 and 162). These correlated residues form a clear physical connectivity through van der Waals interactions between the tetramer interface (Trp binding sites), the dimer interface (Phe binding sites) and the active sites. This suggests that the second cluster of correlated residues may be responsible for the communication between allosteric ligand binding sites and the active sites.

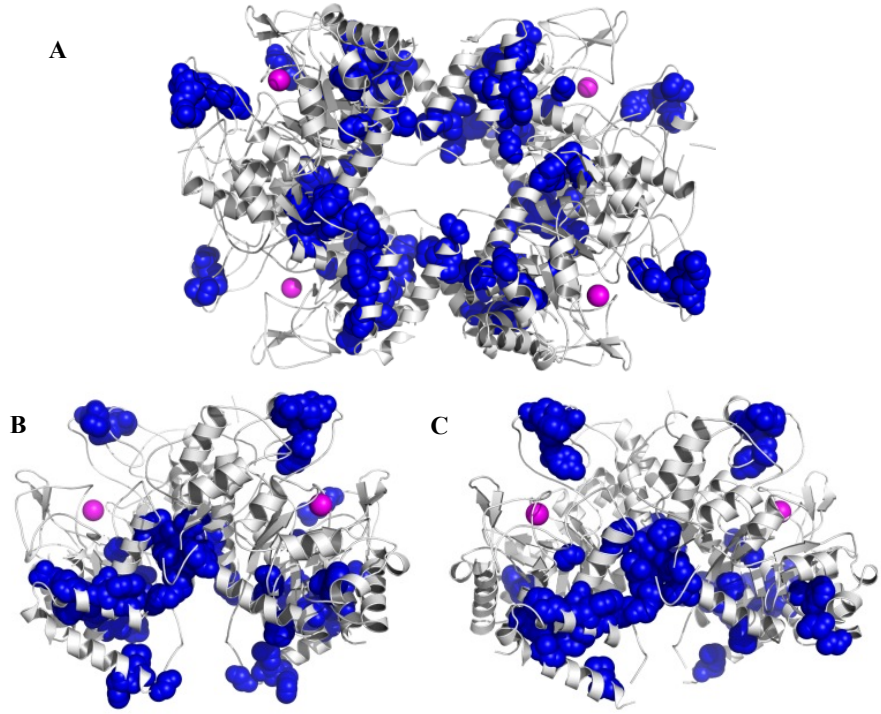


Figure 6. 23. Cluster 2 from SCA·MD analysis of ligand-free *Mtu*DAH7PS. The enzyme is displayed in white cartoon representation; the metal ions Mn^{2+} are shown as magenta spheres to indicate positions of the active sites. The residues found in cluster 2 of ligand-free *Mtu*DAH7PS are displayed as blue spheres on (A) the tetramer, (B) chains A and B, and (C) chains C and D of *Mtu*DAH7PS.

A few residues in cluster 2 are located on helices α2a and α2b which are the two additional helices to the core barrel structure. Superposition of the ligand-free and the Trp+Phe bound crystal structures of *Mtu*DAH7PS also show that there is slight shift in position for these two helices upon ligand binding (discussed in Chapter 4). This suggests that the motion of helices α2a and α2b might play a role in the communication mechanism. Residue 190 (Gly), located at the beginning of helix α2a, was identified in the second cluster in all of the four chains on the tetramer, indicating the movement of this residue is very important for the communication mechanism. However,

Gly190 does not have a side chain functionality that could mediate the signal transmission through side chain conformational fluctuation, so the importance of this residue may be due to its unusual flexibility compared to all other types of amino acids. Interestingly the end of helix α2b is also a Gly residue (Gly232), which was identified in this cluster of correlated residues in chains A and B. With the presence of these two Gly residues, helices α2a and α2b are capable of greater amount of movement compared to other secondary structure elements within the tetramer, and the motion of these two helices may be important for communication between the ligand binding sites and the active site. Therefore, these two residues could be interesting "second-shell" targets for mutagenesis experiments to study the communication mechanism. For example, mutating both G190 and G232 to Proline would reduce the flexibility at these positions, and the importance of the movement of helices α2a and α2b could then be tested.

In summary, SCA·MD analysis was conducted for ligand-free *Mtu*DAH7PS tetramer, and two clusters of correlated residues were identified. The first cluster may be responsible for maintaining relative movement between helices within the same monomer; the second cluster may be responsible for communication across the dimer and tetramer interfaces, between allosteric ligand binding sites and the active sites. The fact that these communication pathways still exist even in the absence of allosteric ligand molecules, suggests that the enzyme is always "listening" for signals from the allosteric ligand binding sites, so the ligand binding event can be "sensed" immediately. This is really an intimate connection between allosteric response and catalytic function.

SCA·MD analysis for Trp+Phe bound *Mtu*DAH7PS. As with the ligand-free system, the SCA·MD procedure and clustering iterations were conducted for Trp+Phe bound *Mtu*DAH7PS tetramer (details of clustering iterations are not shown). The correlation matrix from final iteration of clustering analysis was plotted on a colour scale (Figure 6. 24).

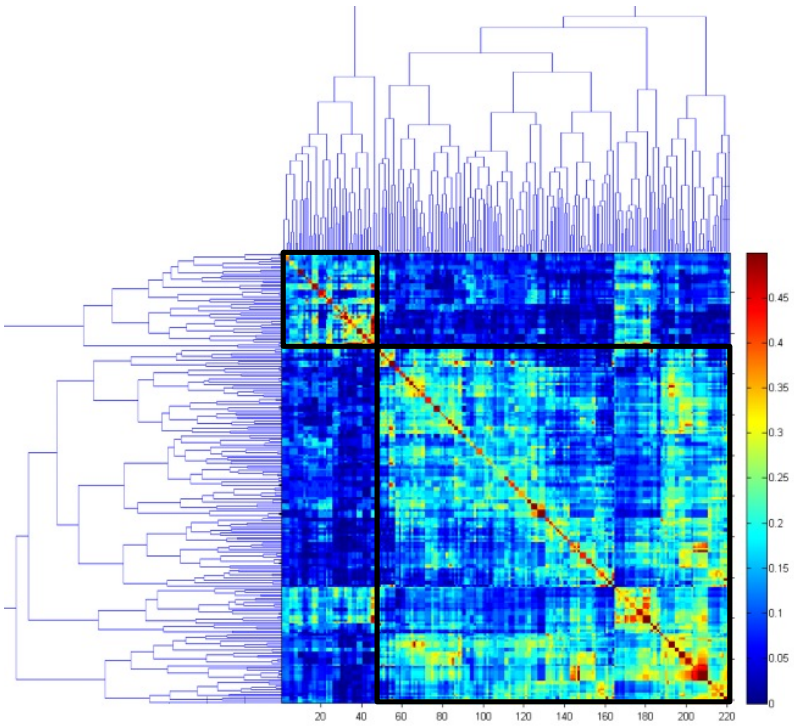


Figure 6. 24. Sorted SCA·MD matrix after the final iteration of clustering analysis for Trp+Phe bound *Mtu*DAH7PS tetramer. The correlation values between pairs of positions are shown using a colour scale, with blue indicating low correlations and red indicating high correlations. The clustering tree was displayed on the top and left of the correlation matrix. The final clusters identified for *Mtu*DAH7PS were highlighted in black boxes.

The SCA·MD analysis for Trp+Phe bound tetramer of *Mtu*DAH7PS identified two final clusters of correlated residues including the following positions:

Cluster 1:

Chain A→
45+361+294+298+297+25+90+271+146+245+131+28
Chain B→
45+361+294+298+297+25+146+245+175+28+40
Chain C→
45+361+294+298+297+25+386+271+446+40
Chain D→
45+361+294+298+297+25+386+90+271+164+245+94+28+40

Cluster 2:

Chain A→
278+61+200+108+454+84+453+192+398+189+218+262+412+57+123+60+186+190+228+143+451+147+145+162+230+435+62+182+197+114+99+58+432+422+184+120+363+185+358+232+65+101+94+175+100+144+157+204+227

Chain B→
278+61+200+108+454+84+453+192+398+189+218+262+412+57+123+60+186+190+228+378+143+144+162+182+435+114+99+432+414+181+173+112+145+147+62+198+120+101+94+100+185+157+204
Chain C→
278+61+200+108+454+84+453+192+398+189+218+262+412+57+123+60+186+190+228+378+451+230+184+120+112+363+162+185+157+358+402+182+232+114+227+179+101+65+100+204+181
Chain D→
278+61+200+108+454+84+453+192+398+189+218+262+412+57+123+60+186+190+228+198+197+378+147+230+435+58+402+358+422+414+184+173+363+451+62+232+181+227+179+175+65

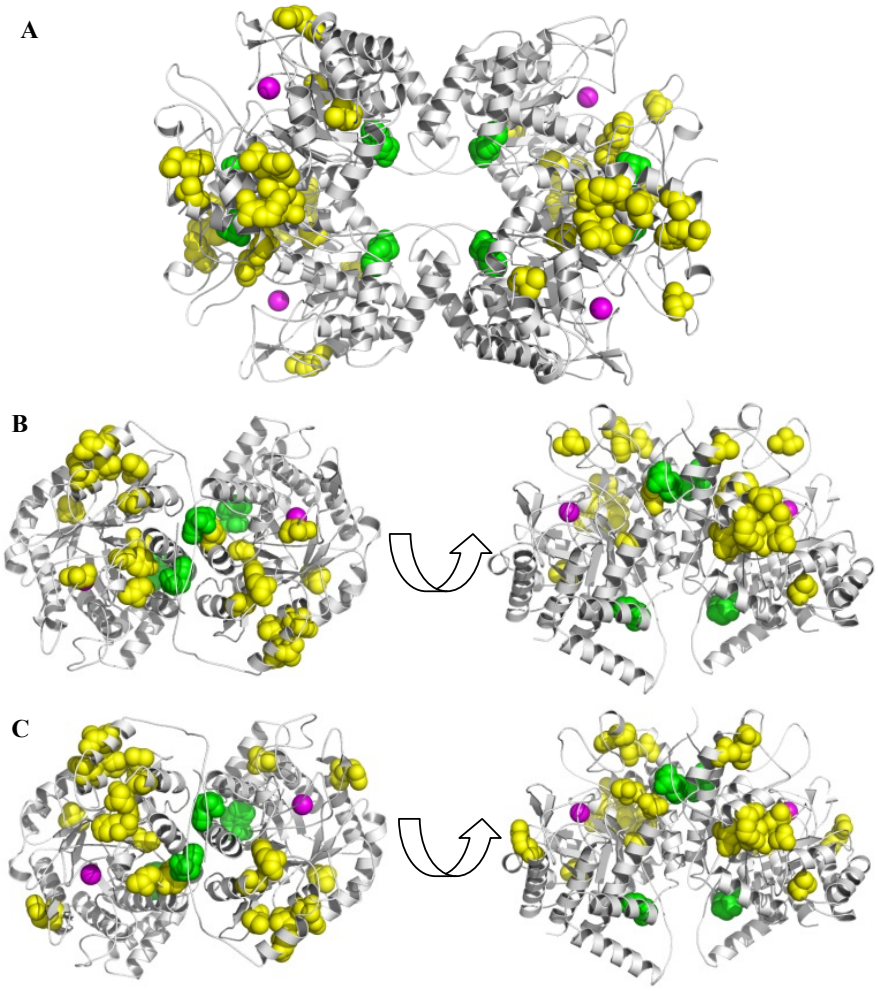


Figure 6. 25. Cluster 1 from SCA·MD analysis of Trp+Phe bound *Mtu*DAH7PS. The enzyme is displayed in white cartoon representation; the bound ligands are shown as green spheres and the metal ions Mn^{2+} are shown as magenta spheres to indicate positions of the active sites. The residues found in cluster 1 of Trp+Phe bound *Mtu*DAH7PS are displayed as yellow spheres on (A) the tetramer, (B) chains A and B, and (C) chains C and D of *Mtu*DAH7PS.

These two clusters of highly correlated residues were then mapped onto the crystal structure of Trp+Phe bound *Mtu*DAH7PS, to examine the locations of

these residues on the tertiary structure of the enzyme, and to investigate the possible physical meaning of these clusters.

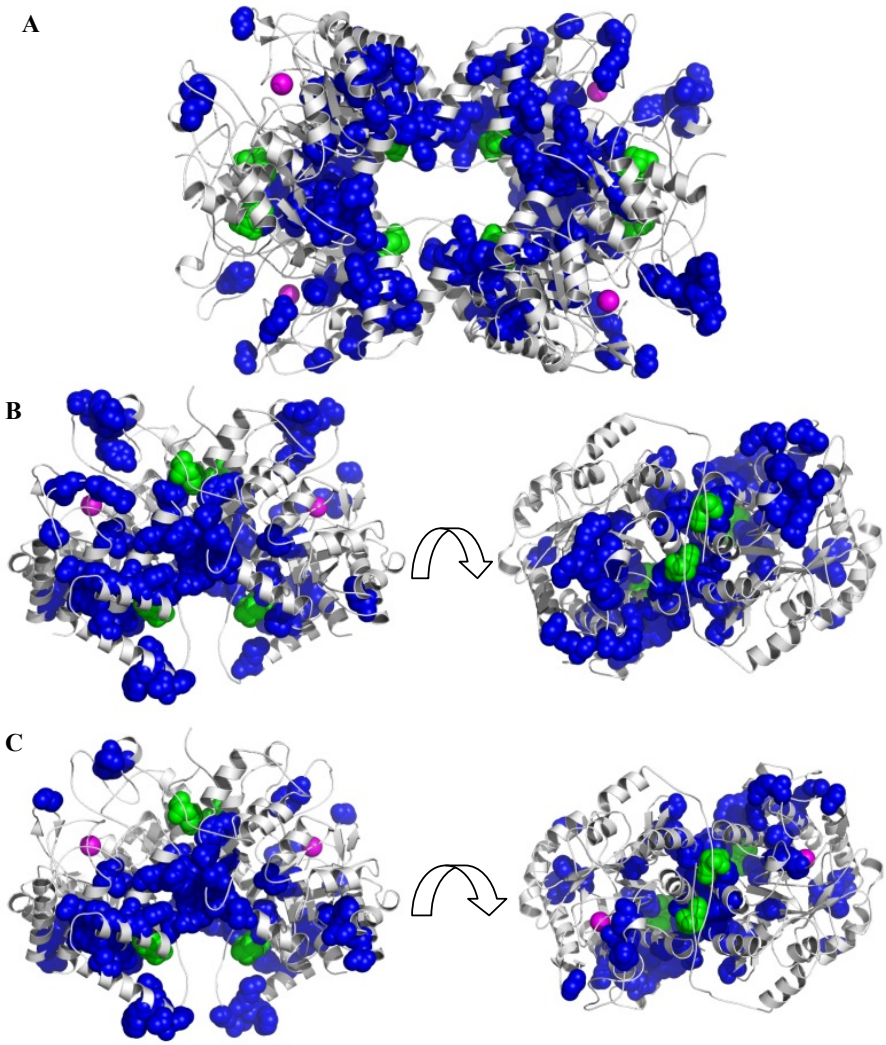


Figure 6. 26. Cluster 2 from SCA·MD analysis of Trp+Phe bound *Mtu*DAH7PS. The enzyme is displayed in white cartoon representation; the bound ligands are shown as green spheres, and the metal ions Mn^{2+} are shown as magenta spheres to indicate positions of the active sites. The residues found in cluster 2 of Trp+Phe *Mtu*DAH7PS are displayed as blue spheres on (A) the tetramer, (B) chains A and B, and (C) chains C and D of *Mtu*DAH7PS.

Cluster 1 of correlated residues in Trp+Phe bound *Mtu*DAH7PS tetramer contains 47 residues (Figure 6. 25). Similar to cluster 1 in the ligand-free system, the residues in this cluster are mostly found between helices within the same monomer. No connectivity was found across the tetramer interface. This cluster contains residues at similar positions as those found in cluster 1 of ligand-free system, e.g. residues 25, 298 and 197. This cluster of residues

may be important for maintaining relative movement between helices within the same monomer, and is unlikely to contribute to the mechanism of allosteric communication.

Cluster 2 of correlated residues in Trp+Phe bound *Mtu*DAH7PS tetramer contained 174 residues (Figure 6. 26). This cluster is similar to the cluster 2 that was found for the ligand-free system, but it is more extended since it contains larger number of residues, and the communication pathway between allosteric binding sites and active sites becomes "smoother" since the residues involved can be well-connected in space through van der Waals interactions. The residues found in this cluster are located mostly around the tetramer interface (e.g. residues 228 and 227), Trp binding sites (e.g. residues 123, 114 and 192), and also the dimer interface (e.g. residues 184, 181 62 and 189), close to Phe binding sites (e.g. residues 94, 175 and 173). Some residues are found on the active site loops (e.g. residues 147, 145 and 143). This suggests that the second cluster of residues is likely to contribute to communication mechanism between the allosteric ligand sites and the active sites when both Trp and Phe are bound.

SCA·MD analysis for Trp-only bound *Mtu*DAH7PS. The same SCA·MD analysis was conducted for the Trp-only bound system (details of clustering iterations are not shown). The correlation matrix from final iteration of clustering analysis was plotted using a colour scale (Figure 6. 27).

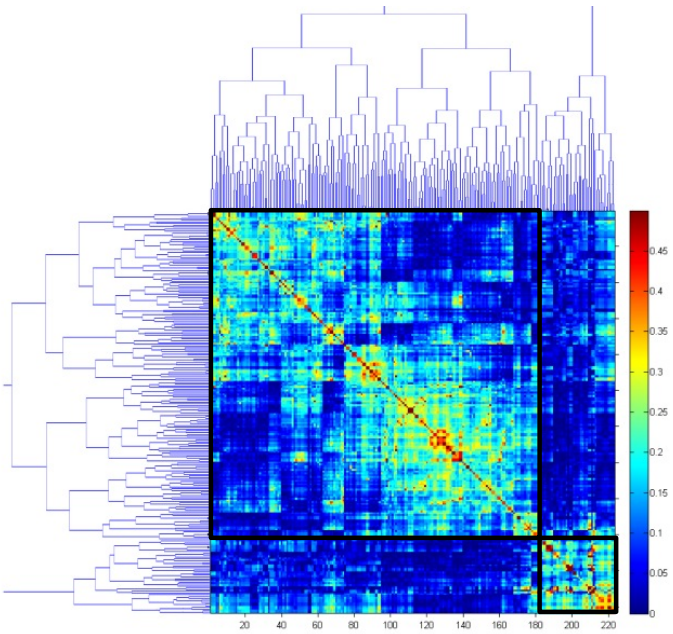


Figure 6. 27. Sorted SCA·MD matrix after the final iteration of clustering analysis for Trp-only bound *Mtu*DAH7PS tetramer. The correlation values between pairs of positions are shown using a colour scale, with blue indicating low correlations and red indicating high correlations. The clustering tree was displayed on the top and left of the correlation matrix. The final clusters identified for *Mtu*DAH7PS were highlighted in black boxes.

The SCA·MD analysis for Trp-only bound tetramer of *Mtu*DAH7PS identified two final clusters of correlated residues including the following positions:

Cluster 1:

Chain A→
245+40+25+28+202+103+105+102+101+175+94+412
Chain B→
245+40+25+294+298+297+271+446+90+175+65+94+101+262
Chain C→
245+40+25+146+103+101+412
Chain D→
245+40+25+294+28+271+65+262

Cluster 2:

Chain A→
228+186+198+454+200+120+190+185+112+451+435+230+108+192+227+197+232+123+114+189+173+204+60+61+145+162+147+57+157+144+62+182+181+84+262+278+90+65+378+398+231+201+143+363+402+221+100
Chain B→
228+186+198+454+200+120+190+185+112+451+435+230+108+192+227+197+232+123+114+189+173+204+60+61+145+162+147+57+157+144+62+

182+181+84+378+358+363+100+218+453+412
Chain C→
228+186+198+454+200+120+190+185+112+451+435+230+108+192+227+197+232+123+114+189+173+204+60+61+145+162+147+57+157+144+62+182+181+84+143+378+218+453+363+414+221+231+110+358+278+65+94+262+175
Chain D→
228+186+198+454+200+120+190+185+112+451+435+230+108+192+227+197+232+123+114+189+173+204+60+61+145+162+147+57+157+144+62+182+181+84+231+110+100+218+453+358+101+90+412+175+94

Same as before, mapping these two clusters of residues onto crystal structure of Trp-only bound *Mtu*DAH7PS tetramer (Figure 6. 28) show that the first cluster identified in the Trp-only system contains similar members to that found in the ligand-free system, and also in the Trp+Phe bound system, e.g. residue 25, 298 and 297. This first cluster contains 41 residues and most of these are located between helical structures within the same monomer, but not across tetramer of dimer interfaces. Therefore, as was suggested in the previously analysed ligand-free and Trp+Phe bound systems, this cluster of residues in the Trp-only system may be important for maintaining relative movement between helices within the same monomer, and is unlikely to contribute to the mechanism of allosteric communication.

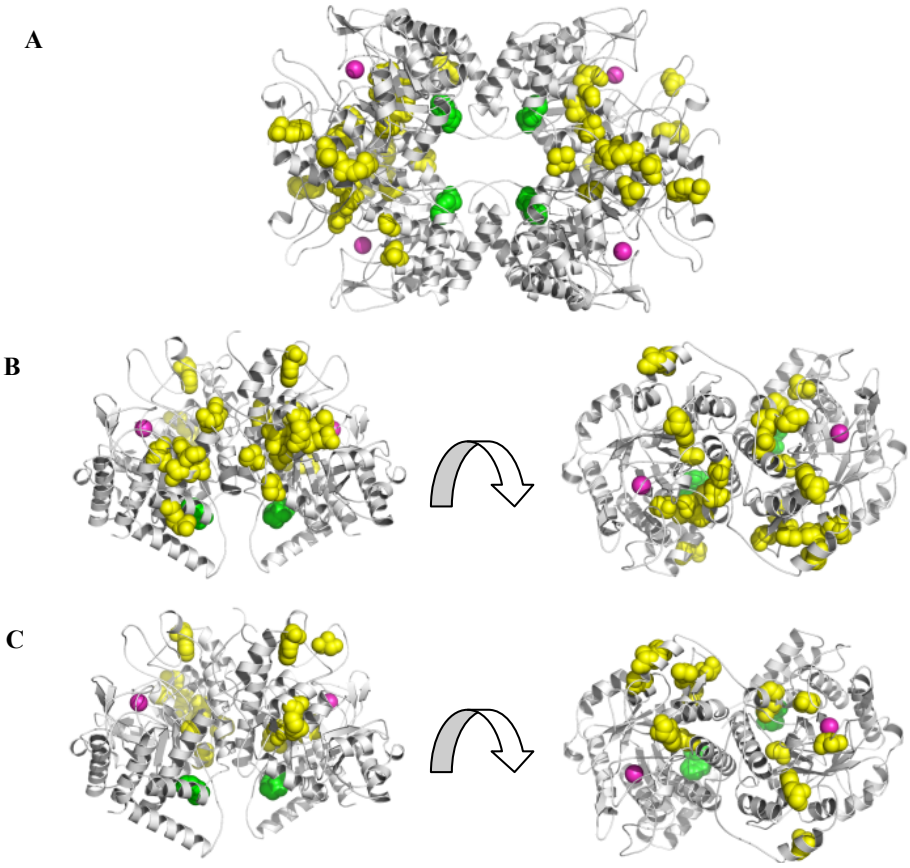


Figure 6. 28. Cluster 1 from SCA·MD analysis of Trp-only bound *Mtu*DAH7PS. The enzyme is displayed in white cartoon representation; the bound ligands are shown as green spheres and the metal ions Mn^{2+} are shown as magenta spheres to indicate positions of the active sites. The residues found in cluster 1 of Trp-only bound *Mtu*DAH7PS are displayed as yellow spheres on (A) tetramer, (B) chains A and B, and (C) chains C and D of *Mtu*DAH7PS.

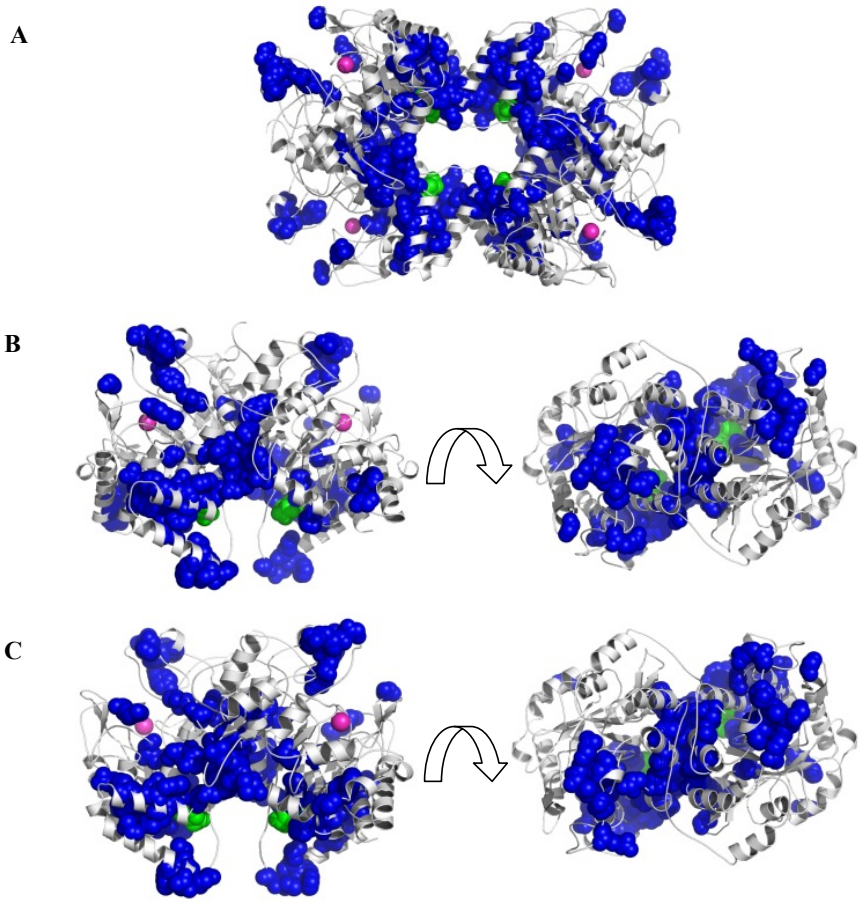


Figure 6. 29. Cluster 2 from SCA·MD analysis of Trp-only bound *Mtu*DAH7PS. The enzyme is displayed in white cartoon representation; the bound ligands are shown as green spheres and the metal ions Mn^{2+} are shown as magenta spheres to indicate positions of the active sites. The residues found in cluster 2 of Trp-only bound *Mtu*DAH7PS are displayed as yellow spheres on (A) the tetramer, (B) chains A and B, and (C) chains C and D of *Mtu*DAH7PS.

Cluster 2 of correlated residues in Trp-only bound *Mtu*DAH7PS tetramer contains 182 residues (Figure 6. 29). This cluster is similar to the cluster 2 that is found Trp+Phe bound system. The residues found in this cluster are located mostly around the tetramer interface (e.g. residues 228 and 227), Trp binding sites (e.g. residues 123, 114 and 192), and also the dimer interface (e.g. residues 182, 181, 62 and 189), close to Phe binding sites (e.g. residues 94, 175 and 173). Some residues are found in the active site loops (e.g. residues 147, 145 and 143). This suggests, as in the unliganded structure, that communication pathways between the allosteric sites and the active sites are still present even with only Trp binding.

SCA·MD analysis for Phe-only bound *Mtu*DAH7PS. The same SCA·MD analysis was conducted for the Phe-only bound system (details of clustering iterations are not shown here). The correlation matrix from final iteration of clustering analysis was plotted using a colour scale (Figure 6. 30).

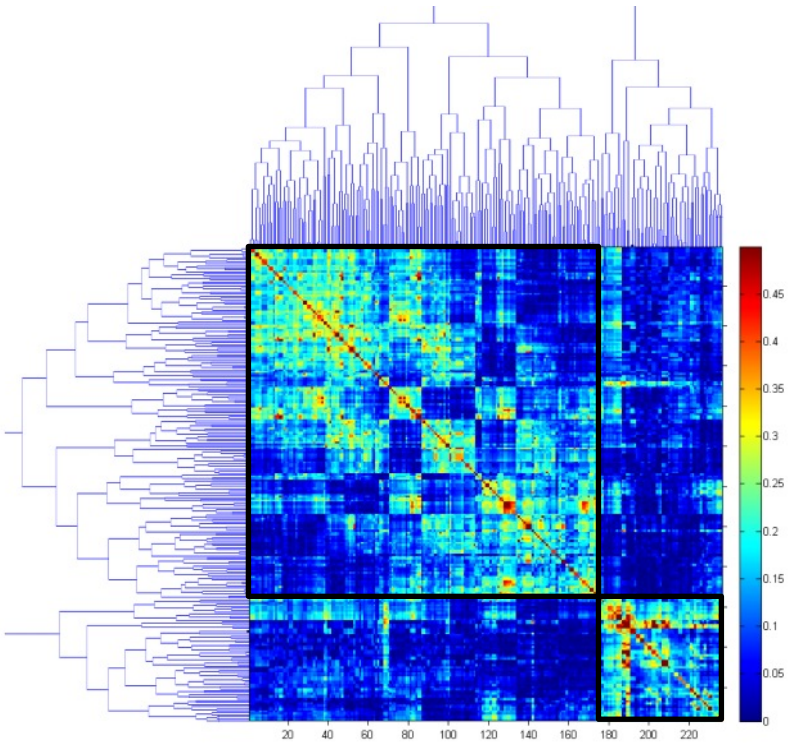


Figure 6. 30. Sorted SCA·MD matrix after the final iteration of clustering analysis for Phe-only bound *Mtu*DAH7PS tetramer. The correlation values between pairs of positions are shown using a colour scale, with blue indicating low correlations and red indicating high correlations. The clustering tree was displayed on the top and left of the correlation matrix. The final clusters identified for *Mtu*DAH7PS were highlighted in black boxes.

The SCA·MD analysis for Phe-only bound tetramer of *Mtu*DAH7PS identified two final clusters of correlated residues including the following positions:

Cluster 1:

Chain A→
25+298+271+100+90+101+175+262+412+131+327+324+297+103+350+386+65
Chain B→
25+298+271+412+131+40+202+386+294+103+297+245+446+101
Chain C→
25+298+271+278+94+28+103+40+105+45+245+90+65
Chain D→

25+298+271+175+412+131+327+324+28+297+40+245+202+105+350+45+294

Cluster 2:

Chain A→
454+84+453+185+57+162+108+204+112+358+144+145+123+189+200+182+230+157+190+227+451+60+186+62+218+114+198+147+120+435+232+61+228+363+181+110+94+378+278+231+192
Chain B→
454+84+453+185+57+162+108+204+112+358+144+145+123+189+200+182+230+157+190+227+451+60+186+62+218+114+198+147+120+435+232+61+228+231+143+378+398+262+402+278+175+100+338+177+99
Chain C→
454+84+453+185+57+162+108+204+112+358+144+145+123+189+200+182+230+157+190+227+451+60+186+62+218+114+198+147+120+435+232+61+228+181+363+175+197+100+262+412+231+58+192
Chain D→
454+84+453+185+57+162+108+204+112+358+144+145+123+189+200+182+230+157+190+227+451+60+186+62+218+114+198+147+120+435+232+61+228+363+391+100+278+414+378+398+181+262+94+143+402+99

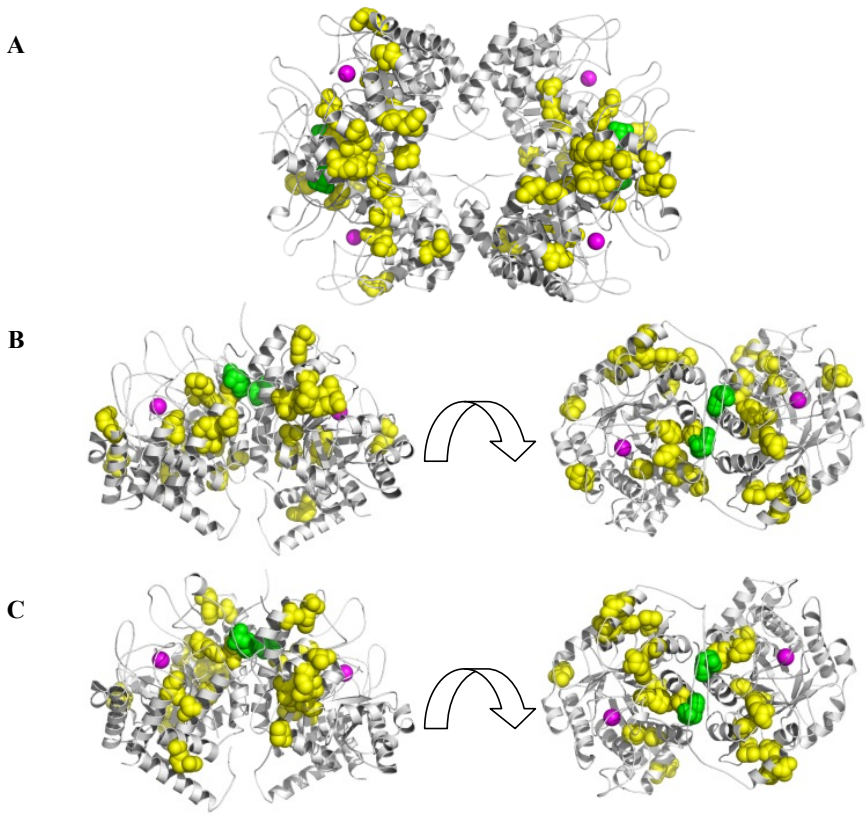


Figure 6. 31. Cluster 1 from SCA·MD analysis of Phe-only bound *Mtu*DAH7PS. The enzyme is displayed in white cartoon representation; the bound ligands are shown as green spheres and the metal ions Mn^{2+} are shown as magenta spheres to indicate positions of the active sites. The residues found in cluster 1 of Phe-only bound

*Mtu*DAH7PS are displayed as yellow spheres on (A) the tetramer, (B) chains A and B, and (C) chains C and D of *Mtu*DAH7PS.

Mapping these clusters of residues onto crystal structure of Phe-only bound *Mtu*DAH7PS tetramer (Figure 6. 31) shows that cluster 1 contains similar members to that found for the ligand-free, the Trp+Phe and the Trp-only bound systems, e.g. residue 25 and 298. This cluster contains 61 residues and most of these are located between helical structures within the same monomer, but not across tetramer of dimer interfaces. Therefore, as in the previously analysed ligand-free, Trp+Phe and Trp-only bound systems, this first cluster of residues in the Phe-only system may be important for maintaining relative movement between helices within the same monomer, and is unlikely to contribute to the mechanism of allosteric communication.

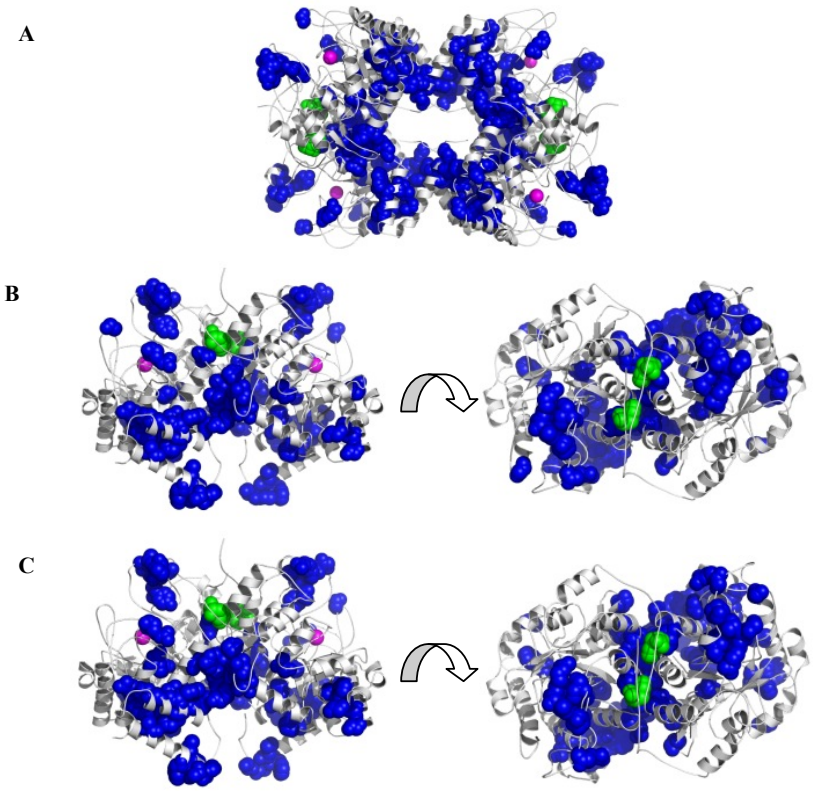


Figure 6. 32. Cluster 2 from MD·SCA analysis of Phe-only bound *Mtu*DAH7PS. The enzyme is displayed in white cartoon representation; the bound ligands are shown as green spheres and the metal ions Mn^{2+} are shown as magenta spheres to indicate positions of the active sites. The residues found in cluster 2 of Phe-only bound *Mtu*DAH7PS are displayed as blue spheres on (A) the tetramer, (B) chains A and B, and (C) chains C and D of *Mtu*DAH7PS

Cluster 2 of correlated residues in Phe-only bound *Mtu*DAH7PS tetramer contains 175 residues (Figure 6. 32). This cluster is similar to the cluster 2 found in Trp+Phe and Trp-only bound systems. The residues found in this cluster are located mostly around the tetramer interface (e.g. residues 228 and 227), Trp binding sites (e.g. residues 123 and 114), and also the dimer interface (e.g. residues 182, 181, 62 and 189), close to Phe binding sites (e.g. residues 94 and 175). Some residues are found in the active site loops (e.g. residues 147 and 145). This suggests that communication pathways between the allosteric sites and the active sites are still present even with only Phe binding.

Mutual residues in cluster 2 of different systems. Since the second cluster identified in both the ligand-free and ligand-bound systems may be important for communication between the allosteric sites and the active site, analysis of the mutual residues in cluster 2 of different systems would give us some idea of the most critical members of this communication pathway and provide a list of potential mutagenesis targets (
Table 6. **2**).

It is interesting that residue Gly190 (discussed in ligand-free system) is identified in Cluster 2 of all four systems, indicating the movement of the two additional helices α2a and α2b is definitely important for communication mechanism. The other Gly residue at the end of helix α2b, Gly232, appears in this cluster for both single ligand-bound systems. Therefore, mutating these two Gly residues to Pro may restrict the potential movement of helices α2a and α2b at the tetramer interface, and in turn may terminate the communications between different sites on the *Mtu*DAH7PS tetramer.

Residue Met228 is another interesting target. This residue forms part of the tetramer interface and interacts with the Met228 from the other monomer. Mutating this residue to Ala may eliminate communications across the tetramer interface, but may also disrupt the formation of the tetrameric complex. Gln108 and Phe454 are both located on the main barrel structure with their side chains pointing towards the two additional helices α2a and

α2b, while Trp200 is located on helix α2a facing the main barrel. Mutating these three residues to Ala may stop the signal transmission from α2a and α2b to the main barrel and thus to the active sites.

Table 6. 2. Mutual residues in cluster 2 of different systems.

Mutual to all systems	**Mutual to the three ligand-bound systems**	**Mutual to only single ligand systems**	**Mutual to only single ligand and ligand-free systems**
Met228 Phe454 Trp200 Gly190 Gln108 Pro57	Leu186 Lys123 Ser189 Val60 Pro61 Gly84	His198 Pro120 Ala185 Val112 Glu451 Arg435 Ala230 Phe227 Gly232 Thr114 Phe204 Tyr147 Leu144 Ser62 Leu182	Arg145 Arg162 Pro157

The residues mutual to all three ligand-bound systems are mostly located on the dimer interfaces (e.g. residues 186, 189, 60 and 61), suggesting they may be the critical residues responsible for communication across the dimer, or from Trp binding sites to Phe binding sites, upon ligand binding. However, mutating these residues needs to proceed with caution, as they may disrupt the dimer interface, and may result in total loss of function.

Those residues mutual to only single ligand systems are located in the Trp binding sites (e.g. residues 114 and 230), in the tetramer interface (e.g residues 227, 120), between the main barrel and the additional helices α2a and α2b (e.g. residues 451, 112 and 204), on the dimer interface (e.g. residues 182, 185 and 62). Mutation experiments can also be designed around these targets.

6.4.3 Summary

SCA·MD analysis was conducted for ligand-free and the three ligand-bound *Mtu*DAH7PS tetramer systems (Trp+Phe, Trp-only and Phe-only bound). In each of the four systems, two clusters of highly correlated residues were identified. The two clusters show similarity across the four different systems. The smaller cluster (cluster 1 in each system) generally contains residues between helices within the same monomer, and has no connectivity across the tetramer interfaces or dimer interfaces. Therefore it is unlikely for this cluster to contribute to the communication mechanisms between different allosteric sites and the active sites. Although members of this cluster may change in each system, their general locations remained the same.

The larger cluster (cluster 2 in each system) generally contains residues at the tetramer interfaces, Trp binding sites, dimer interfaces, Phe binding sites and the active sites. Therefore it is considered to be the dynamic pathway responsible for communications between allosteric sites and the active sites within the tetramer structure of *Mtu*DAH7PS. The fact that this pathway is still present in the ligand-free system suggests that it is always "functional". Even in the absence of ligand molecules, this pathway can be responsible for communication and cooperatively between different monomers, and at the same time it is listening for signals of ligand binding at those allosteric sites. The pathway formed by this cluster of residues is more extended and better connected through van der Waals interactions in space in the ligand-bound systems compared to the ligand-free system. Binding of single ligand is enough to trigger this extension of pathway, suggesting that the pathway is responsible for the synergy between the binding events of different ligands.

Several interesting residues were identified as mutagenesis targets to stop the communication between different allosteric sites and between allosteric sites and the active sites. Among these the most attractive targets are Gly190 and Gly232.

6.5 SCA·MD analysis for G190P/G232P mutant

The two Gly residues (190 and 232) are located at the beginning and the end of helices α2a and α2b. These two residues, especially Gly190, were identified in cluster 2 of important residues from SCA·MD analysis, and were proposed to be important for signal transmission between the monomers in the tetramer structure of *Mtu*DAH7PS and communication between Trp/Phe binding sites and the active site. *In silico* mutations were done for this particular double mutant and 20 ns MD simulation was conducted on this double mutant to examine whether there is any change in molecular motion caused by the mutation and whether the communication of Trp binding sites with the Phe binding sites and the active sites is affected.

6.5.1 In silico *mutations and set up of MD simulation (20 ns)*

To prepare the starting structure of the G190P/G232P double mutant, the last frame of the 20 ns MD simulation for ligand-free wild type *Mtu*DAH7PS was taken, including water molecules in the water box and also the ions that were added to balance the net charge of the system. The *in silico* mutations were then conducted using the "mutate residue" function in VMD (*118*), to change Gly190 and Gly232 in all of the four monomers to Pro residues. The structure is then ready for MD simulation because water molecules have already been added and the net charge of the system is already balanced when preparing the system for the ligand-free wild type simulation. All other simulation setups are the same as in the wild type systems. The molecular dynamics for the double G190P/G232P mutant was conducted for 20 ns at 2 fs time steps using NAMD (*36*).

Protein backbone RMSD values over 20 ns simulation period compared to ligand-free wild type *Mtu*DAH7PS. The backbone RMSD values for G190P/G232P mutant over the 20 ns of MD simulation were plotted along the time scale together with the RMSD values from the ligand-free wild type *Mtu*DAH7PS (Figure 6. 33). The plot shows that similar to the wild type

ligand-free system, the simulation of the double G190P/G232P mutant started to equilibrate after ~2.3 ns. The average structure of G190P/G232P mutant during the equilibrated time period (2.3 ns to 20 ns) was compared to the average structure from the ligand-free wild type simulation, and these two average structures match with each other with RMSD value of 1.540 Å (matching 1728 C^{α} atoms). This suggests that the double mutation of Gly to Pro did not cause any major change in the overall conformation of *Mtu*DAH7PS tetramer.

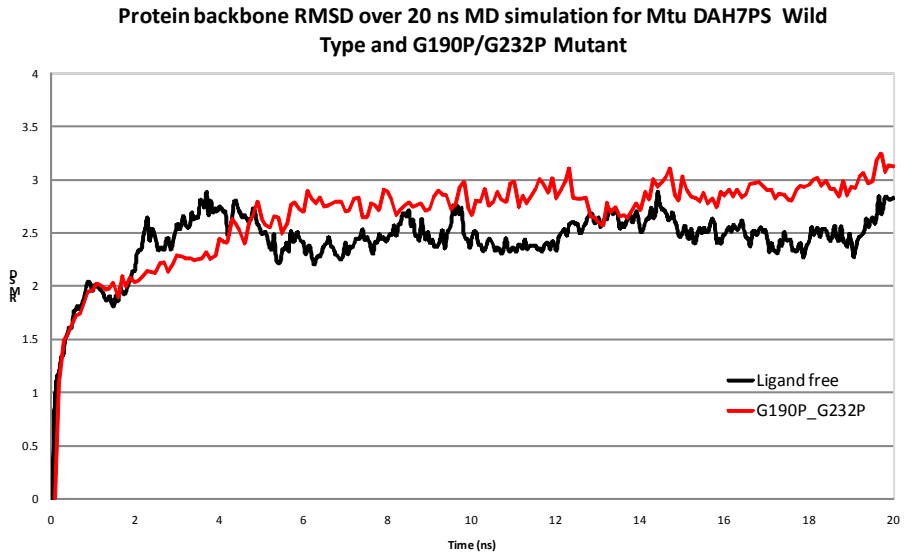


Figure 6. 33. Average backbone RMSD values of wild type ligand-free and G190P/G232P double mutant of *Mtu*DAH7PS during 20 ns MD simulations. The RMSD values were calculated using the first frame of the trajectory as the reference structure.

Chain-averaged residue RMSF values over 20 ns of MD simulation compared to ligand-free wild type *Mtu*DAH7PS. Although the double mutations did not greatly affect the average conformation of the *Mtu*DAH7PS tetramer, it did result in changes in the flexibility of specific regions of the structure. The average RMSF values of individual amino acids of *Mtu*DAH7PS were calculated, in order to examine the changes in dynamic properties caused by the double G190P/G232P mutation. For this analysis, trajectory frames collected from the equilibrated simulation period were used (after 2.3 ns), and all non-hydrogen atoms of each amino acid (backbone and side-chain atoms) were included in the calculation. The chain averaged residue RMSF values from the G190P/G232P mutant system is

plotted together with that from the ligand-free wild type system (Figure 6. 34). In both systems, the residues on the loop regions gave greater RMSF values (more flexible) and the residues on secondary structures such as helices and sheets showed relatively smaller RMSF values (less flexible).

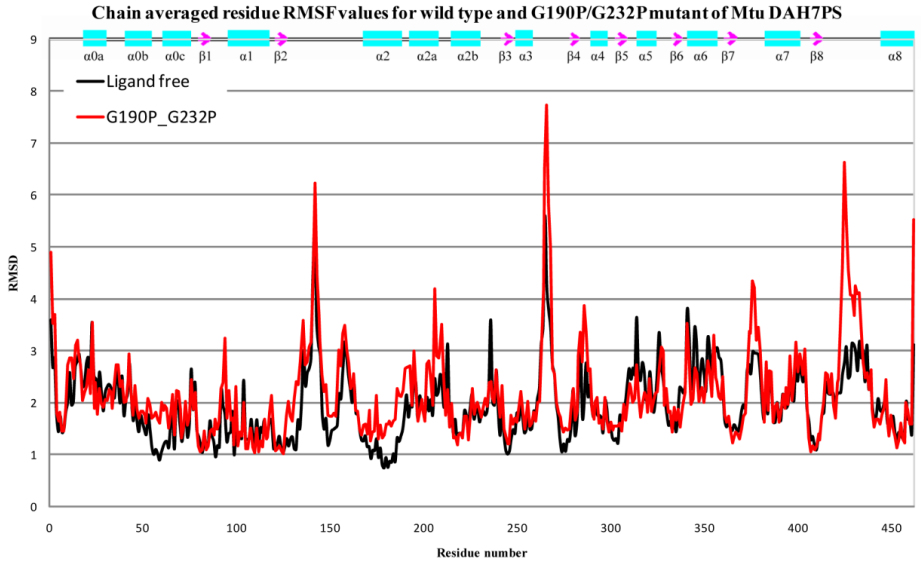


Figure 6. 34. Chain-averaged RMSF values residue-by-residue for the double G190P/G232P mutant and ligand-free wild type of *Mtu*DAH7PS complexes.

Comparison of the RMSF values between the mutant and the wild type systems shows that the double mutation of G190P/G232P causes increase in flexibility around helix α2 (Figure 6. 35), which makes up the tight dimer interface (Figure 6. 36). Most residues in the Phe binding sites become more flexible (Figure 6. 35), as a result the binding of Phe molecules to this mutant might be disfavoured due to increased entropy cost upon binding. The flexibility of the loop β2α2 in the active site is also increased (Figure 6. 35, Figure 6. 36), and this suggests that the double mutation might inhibit the binding of substrates due to unfavourable entropy, similar to that found in the Trp+Phe bound system. However, the flexibility of residues on the helices α2a and α2b does not decrease as intended with the G190P/G232P mutations. Most of the residues on helix α2a show increase in flexibility with the double G/P mutation, and only the residues on the α2b helix show slight decrease in flexibility. Several residues in the Trp binding sites become less flexible with the double mutation, including residues 107, 100, 111, 114, 121 and 123. This suggests that the mutation may favour Trp binding, due to less entropy

cost upon binding. Another possible effect of the mutations is that they might partially mimic the Trp binding signal at the Trp binding sites, because Trp binding at the wild type ligand-free *Mtu*DAH7PS will cause the residues surrounding it to be less flexible, which is partially observed in the double mutant system.

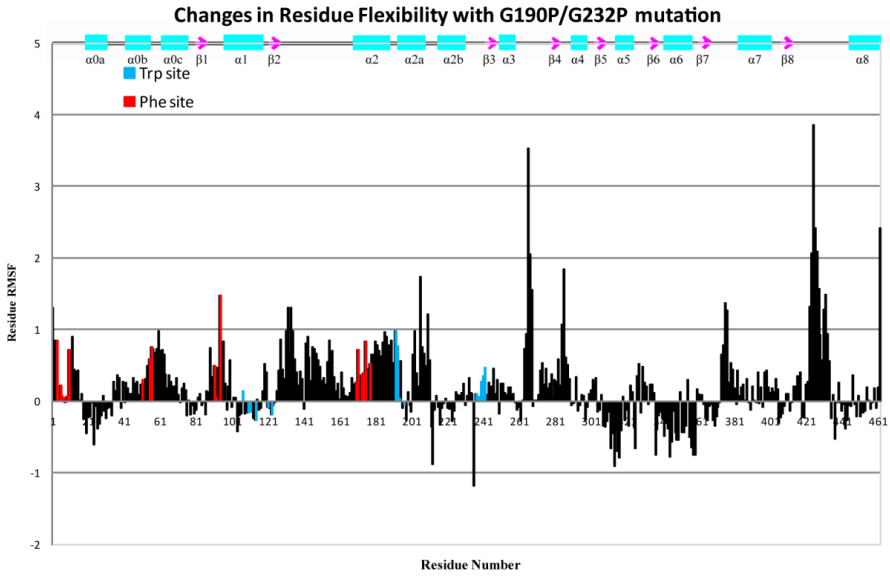


Figure 6. 35. Changes in residue flexibility caused by G190P/G232P mutations. Positive values indicate increases in flexibility compared to the ligand-free wild type system, and negative values indicate decreases in flexibility. The residues involved in Trp and Phe binding sites are highlighted with different colours.

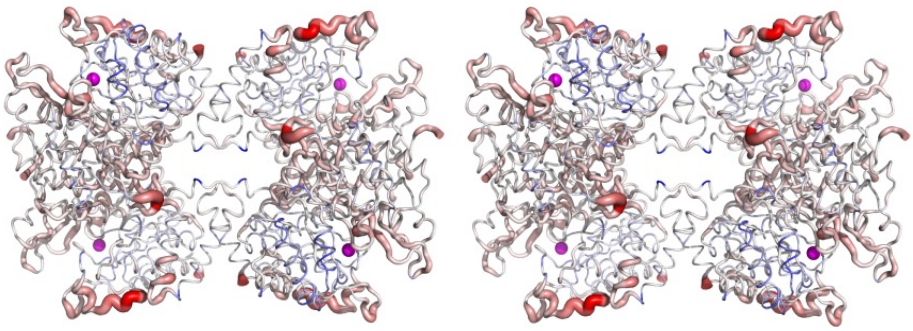


Figure 6. 36. Changes in residue flexibility (ΔRMSF) in *Mtu*DAH7PS caused by the double mutation of G190P/G232P as revealed by molecular dynamics simulations.

The loops β7α7 and β8α8 also show large increase in flexibility with the double G/P mutations (Figure 6. 35). These loops are located along the side of the tetramer complex near the active site (Figure 6. 36). These residues may not directly contribute to substrate binding; however, they may play a role in the formation of the complex between *Mtu*DAH7PS and chorismate

mutase (*77*). As seen in the crystal structure (Figure 6. 37), the loops β7α7 and β8α8 are located close in space to the chorismate mutase when it is in complex with *Mtu*DAH7PS. Increase of flexibility in these loops may diminish the enzyme's affinity for chorismate mutase.

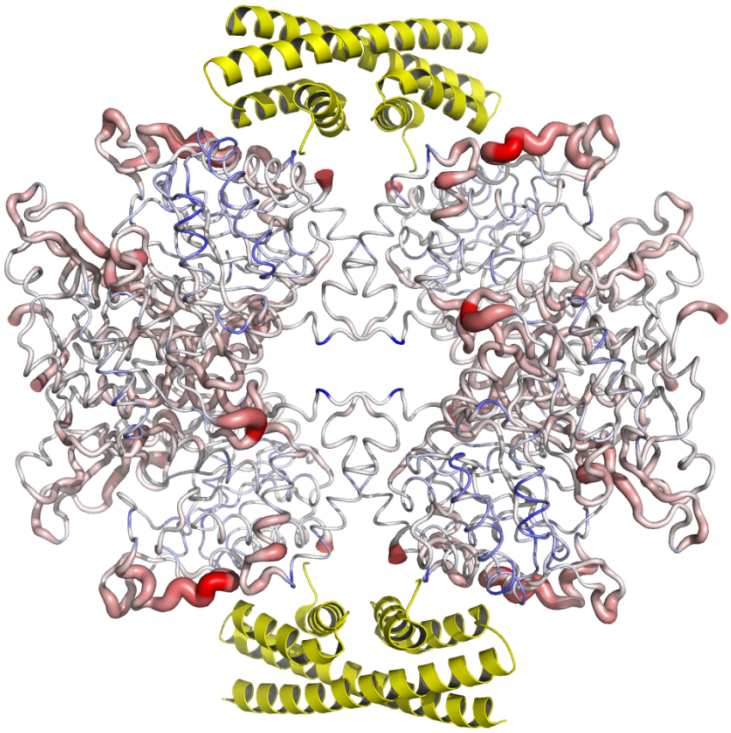


Figure 6. 37. *Mtu*DAH7PS in complex with chorismate mutase (coloured in yellow). The changes in residue flexibility caused by G190P/G232P double mutations are mapped on to the *Mtu*DAH7PS tetramer, with red indicating increase in flexibility and blue indicating decrease in flexibility.

6.5.2 SCA·MD analysis for G190P/G232P double mutant of Mtu*DAH7PS*

In order to examine the correlation of movements between residues on the mutant tetramer and compare the residues involved in the communication network to the ligand-free wild type system, the SCA·MD analysis was conducted for the double G190P/G232P mutant system with the same clustering procedures as used for the wild type systems (details of clustering iterations are not shown). The correlation matrix from final iteration of clustering analysis was plotted using a colour scale (Figure 6. 38).

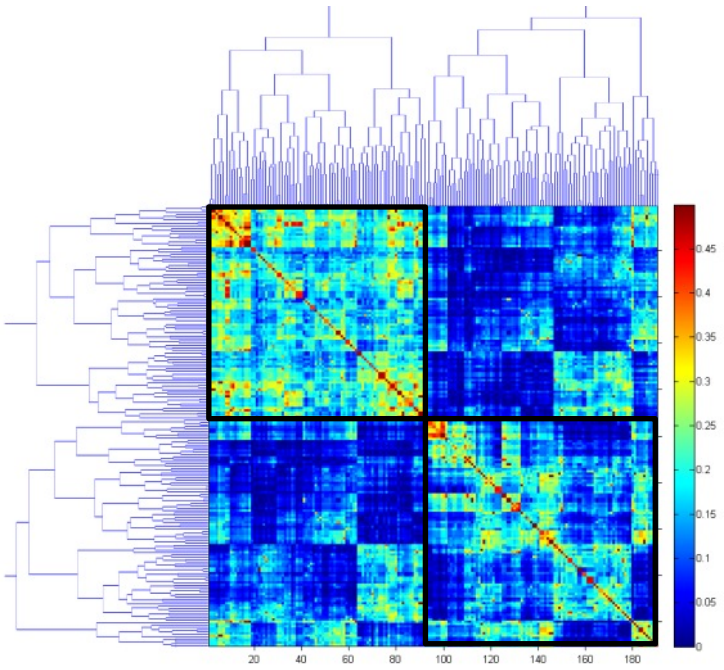


Figure 6. 38. Sorted SCA·MD matrix after the final iteration of clustering analysis for G190P/G232P double mutant of *Mtu*DAH7PS. The correlation values between pairs of positions are shown using a colour scale, with blue indicating low correlations and red indicating high correlations. The clustering tree was displayed on the top and left of the correlation matrix. The final clusters identified for *Mtu*DAH7PS were highlighted in black boxes.

Two clusters of highly correlated residues were identified in the double G190P/G232P mutant system, and the two clusters include the following residues:

Cluster 1

Chain A→
182+185+57+123+198+181+204+114+61+60+457+391+435+100+62+120+278+110+363+453+451+232+84+454+227+186+228+200
Chain B→
182+185+57+123+198+181+114+186+189+227+190+192+143+120+84+144+112+162
Chain C→
182+185+57+123+198+181+189+200+186+457+452+231+230+190+192+197+451+100+112+358+120+108+204+454+227+228
Chain D→
182+185+57+123+198+181+204+200+62+60+189+190+192+218+197+108+232+61+143+228

Cluster 2:

Chain A→

94+412+262+147+157+145+190+144+162+108+197+173+189+230+112+358
Chain B→
94+412+262+147+157+145+90+65+454+358+378+278+61+62+435+108+60+230+231+110+173+451+453+398+363+228+232+100+204
Chain C→
94+412+262+147+157+145+65+446+90+278+378+61+60+363+143+232+162+144+84+435+218+110+173+453+114
Chain D→
94+412+262+147+157+145+65+175+90+101+100+358+363+378+278+414+453+451+435+162+230+231+120+227+144+186+112+84

The first cluster contains 92 residues in the double G190P/G232P mutant of *Mtu*DAH7PS. Most of these residues are found between the tetramer interface and the dimer interface (Figure 6. 39). They can be observed to form a communication pathway from the Trp binding sites at the tetramer interface (e.g. residues 123, 114, 227 and 228) to the dimer interface near Phe binding sites (e.g. residues 181, 182, 185 and 57). This communication pathway goes around the main barrel structure of each monomer instead of passing through the centre of the barrel (Figure 6. 39B and C). This cluster is similar to cluster 2 in other wild type systems (with or without ligand-bound), except that there are not as many residues found in the active site as in the other wild type systems. MD simulation suggests that the communications between the dimer and tetramer interfaces and between Trp and Phe binding sites are still present. The double mutation of G190P/G232P does not appear to eliminate the allosteric communication pathway as intended.

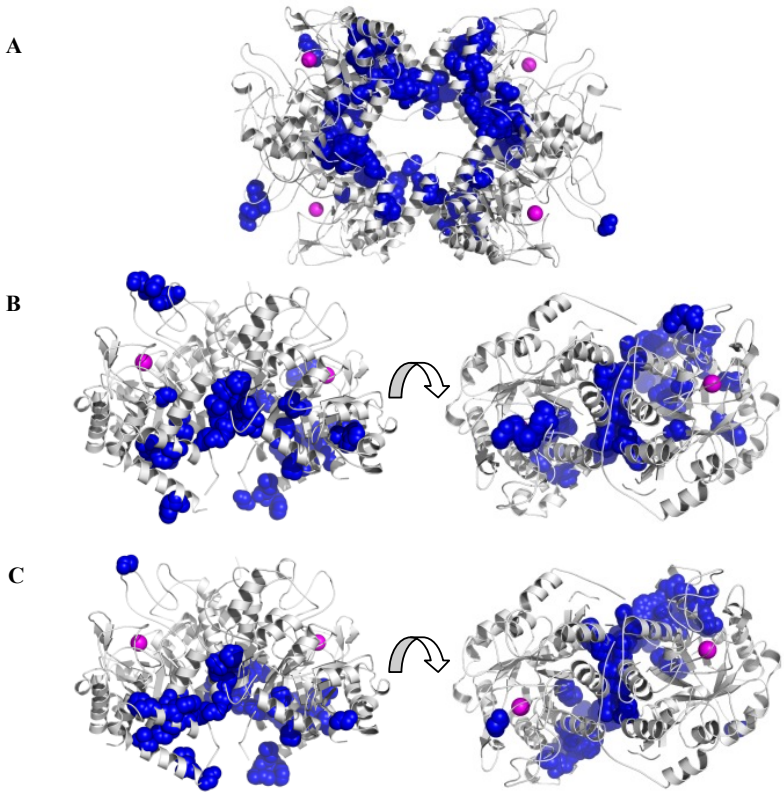


Figure 6. 39. Cluster 1 of correlated residues from SCA•MD analysis for G190P/G232P double mutant of *Mtu*DAH7PS. The enzyme is shown with white cartoon representation, and the active site metal ions Mn^{2+} are shown as magenta spheres to indicate locations of the active site. The correlated residues in cluster 1 are shown as blue spheres on (A) the tetramer, (B) chains A and B, and (C) chains C and D.

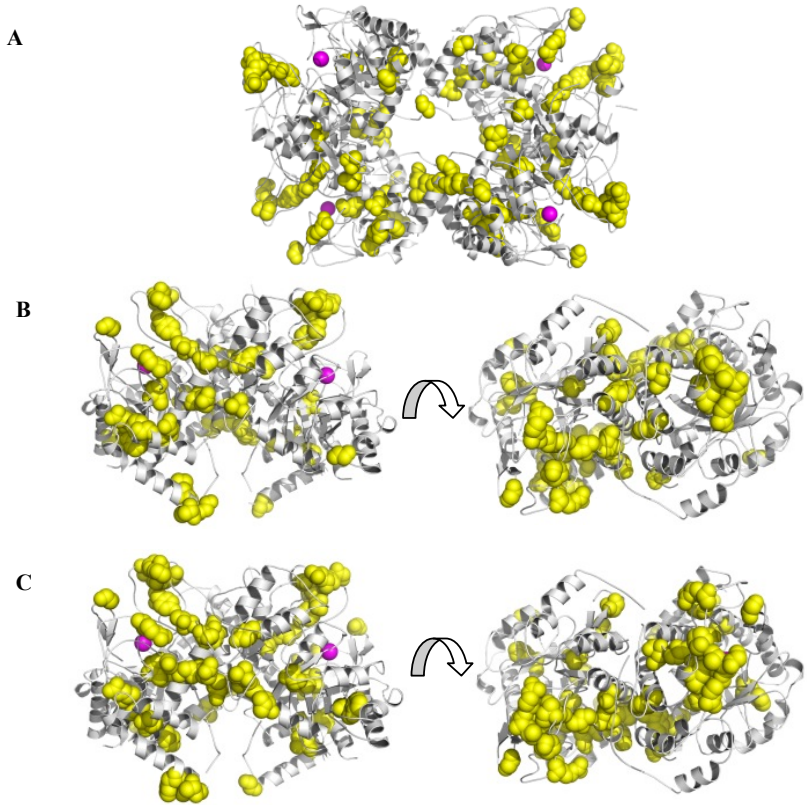


Figure 6. 40. Cluster 2 of correlated residues from SCA•MD analysis for G190P/G232P double mutant of *Mtu*DAH7PS. The enzyme is shown with white cartoon representation, and the active site metal ions Mn^{2+} are shown as magenta spheres to indicate locations of the active site. The correlated residues in cluster 2 are shown as blue spheres on (A) the tetramer, (B) chains A and B, and (C) chains C and D

Cluster 2 contains 98 highly correlated residues in the G190P/G232P double mutant system (Figure 6. 40). The residues in this cluster can be found in the tetramer interface and Trp binding sites (e.g. residues 230, 114, and 227), dimer interface and Phe binding sites (e.g. residues 173, 175, 94, 60 and 186), and the active sites, especially on loop β2α2 in the active site (e.g. residues 145, 147 and 157). The residues in cluster 2 can be observed to form a communication pathway between the dimer interfaces (Phe binding sites) and the active site, and (to some extend) the Trp binding sites at the tetramer interfaces.

In contrast to the wild type system, where only one cluster of residues (cluster 2) is found to be responsible for the allosteric communication between allosteric sites and the active sites, the G190P/G232P double mutant system shows both of the clusters of residues may be capable of transmitting signals between the binding sites. The differences in physical significance

and function between the two clusters found in the mutant system cannot be discriminated easily as in the wild type systems.

6.5.3 Proposals for the experimental verification of MD simulations

Although MD shows that the double G190P/G232P mutations did not eliminate the pathway of allosteric communications between binding sites, the simulation still shows some interesting properties of this mutant system that can be tested using experimental techniques.

1. *Reaction kinetics experiments.* One of the interesting aspects of this mutant system is the effect of mutation on the active site loops. Chain averaged RMSF analysis showed that the active loop β2α2 that is responsible for binding of the substrate E4P, becomes more flexible with this mutation, which is similarly observed in the Trp+Phe bound system. Therefore, this mutation may inhibit the catalytic function of *Mtu*DAH7PS, by disfavouring the binding of E4P. Experiments can be done to measure the reaction kinetics and binding constants of E4P in order to confirm this.

2. *Inhibition studies upon ligand binding.* If the experiment above (reaction kinetics) showed that the mutant is still catalytically active, then it would be interesting to test whether the binding of Trp and Phe, or single amino acids will inhibit the mutant.

3. *ITC experiments for single ligands.* Another good experiment is to test the response of this mutant to the binding of Trp and Phe (and maybe Tyr). This is because MD simulation shows that the double mutation caused the residues at the dimer interface and Phe binding sites to become more flexible, and as a result may disfavour the binding of Phe. However, part of the Trp binding sites become less flexible, this may favour the binding of Trp molecules to the enzyme. Measurement of binding affinities for Trp and Phe can test this prediction.

4. *ITC experiments for combinations of ligands.* MD simulation shows that the allosteric communications still exist in the mutant system. This can be tested by measure binding affinities of Trp alone, and compare to that when Phe is present in the background. The same can be done for Phe molecules. If communications exist between the allosteric sites, then binding of one allosteric ligand may affect the binding affinity of the other.

5. *ITC experiments for MtuDAH7PS in complex with chorismate mutase.* From the analysis of residue RMSF values, MD simulation suggests that the loops β7α7 and β8α8 become more flexible in the mutant system. This may disfavour the binding of chorismate mutase to *Mtu*DAH7PS because the entropy cost upon forming the complex is increased (Figure 6. 37). Measurements of binding affinities for chorismate mutase in the mutant system in comparison with the wild type system may verify this prediction.

6.5.4 Summary

MD simulation (20 ns) was set up for the double mutant G190P/G232P of *Mtu*DAH7PS. The trajectory was analysed using residue RMSF in comparison with the ligand-free wild type system, and SCA·MD analysis to identify the communication networks in the mutant system. Results suggest that the mutant may cause changes in flexibility in the dimer interface, part of the Trp binding site, and also the active sites. Changes in flexibility of loops β7α7 and β8α8 are also predicted, which may affect the formation of complex with chorismate mutase. SCA·MD analysis reveals that the allosteric communication between ligand binding sites and the active sites may still be present in the mutant system, and the mutation is not predicted to work as initially proposed. Several different experiments are proposed that may help to verify predictions made from the MD simulations.

Chapter 7: Summary

Protein molecules are dynamic in nature. The flexibilities within a protein molecule give rise to ensembles of slightly different protein conformations that constantly interconvert via dynamic motions. The importance of protein dynamics to function and regulation has been largely recognized recently (*94-98, 100, 101*). In this thesis, the dynamic behaviour of different enzyme systems was investigated by the application of different molecular modelling techniques.

Part A of this thesis describes a computational study on the dynamic conformational preference of a small peptide CP1B (20-mer), which is a calpain specific inhibitor peptide. Both MCSD and MD simulations were employed, and results showed that in the absence of the enzyme calpain, despite the large flexibility and lack of secondary structures within the peptide, there is a region (residue 10 to 13) that prefers a folded loop like conformation similar to what is known to bind in the active site of calpain. By adopting this folded loop like conformation, the peptide is prearranged to the binding mode, thus decrease the entropy cost upon binding to the enzyme calpain. This conformational preference of the loop region is shown to be one of the significant contributing factors to the inhibition potency of the peptide CP1B.

Part B of this thesis describes the computational study for a large enzyme system (DAH7PS) from *M. tuberculosis*. This tetrameric enzyme exhibits an unusual synergistic allostery by combinations of aromatic amino acids (Trp+Phe, or Trp+Tyr). The binding sites of different allosteric inhibitors (Trp, Phe and Tyr) were revealed by crystallography. Structures of the ligand-bound and ligand-free enzyme indicate that there is no change in gross conformation of the enzyme upon ligand binding, which suggests a change in protein dynamics as the mechanism of inhibition. MD simulations for both the ligand-free and the three ligand-bound systems were conducted in order to identify dynamic changes upon ligand binding. Simulation results showed

that binding of both Trp+Phe results in an increase in flexibility of the E4P binding loop in the active site. The increase in flexibility will increase the unfavourable entropy cost for E4P binding, thus achieves inhibition. MD simulation results also revealed that when *Mtu*DAH7PS is bound with only Trp or only Phe molecules, the dynamics of the binding sites of the other ligand is altered, and the entropy for subsequent ligand binding event is therefore decreased. The residues involved in the communication pathway between the ligand binding sites and the active sites were identified, by combining the MD results with that from the SCA method. Based on these results several targets were proposed for future mutagenesis studies.

The work in this thesis illustrated how a range of molecular modelling techniques can be combined with experimental findings to investigate the molecular basis of enzymic actions. Different modelling methods can be used depending on the questions under investigation. For the small peptide CP1B, due to its flexibility, it is important to have more efficient sampling of the conformational space in order to fully investigate the dynamic conformational preference of this peptide. Therefore, the faster stochastic dynamics simulation with enhanced sampling method (Monte Carlo) was applied. In contrast, for the tetrameric enzyme *Mtu*DAH7PS, only slight changes in conformation and dynamic changes over short time period need to be explored, so the method full atom molecular dynamics simulation was employed.

Computation is a powerful tool as it allows understanding and interpretation of properties of single molecules, allowing intermediate states and transitional conformations to be visualized. The combination of computational techniques and experiments therefore becomes a robust approach that can explain observations from experiments at atomic resolution, predict outcomes from designed experiments, and provide guidance to future research directions.

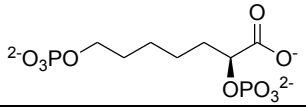
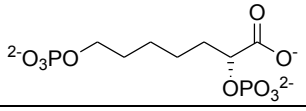
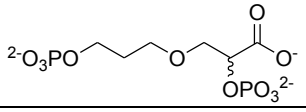
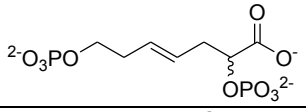
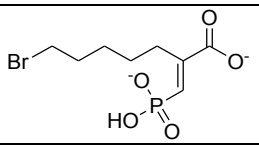
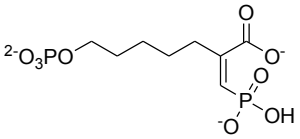
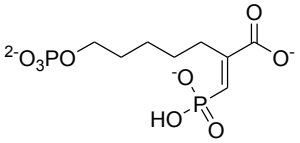
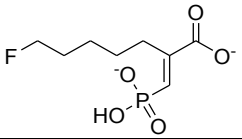
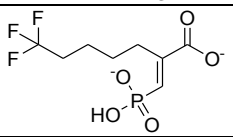
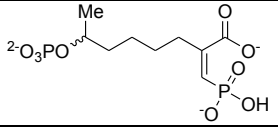
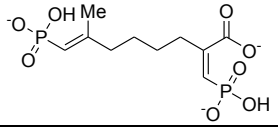
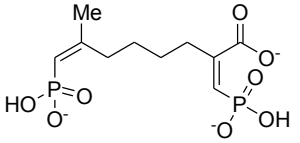
Appendices

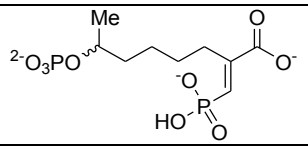
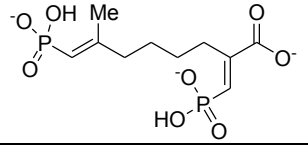
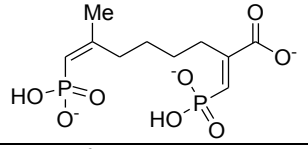
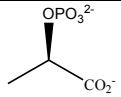
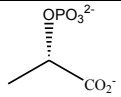
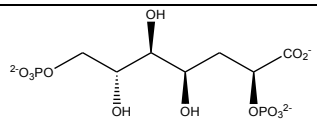
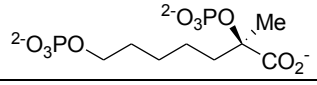
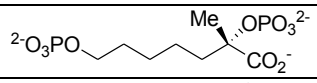
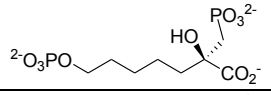
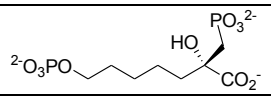
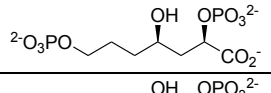
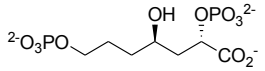
Appendix 1: Sample Configuration script for setting up molecular dynamics simulations with NAMD

```
#############################################################
##ADJUSTABLE PARAMETERS##
#############################################################
structure                mcsd_wb.psf
coordinates              mcsd_wb.pdb
outputName                 mcsd_wb_eq_5ns
set temperature       310
firsttimestep         0
#############################################################
## SIMULATION PARAMETERS##
#############################################################
# Input
paraTypeCharmm             on
parameters
/hpc/home/wji20/cp1b/5nsMD/par_all27_prot_lipid.prm
temperature              $temperature
# Periodic Boundary conditions
# NOTE: Do not set the periodic cell basis if you have also
# specified an .xsc restart file!
if {1} {
cellBasisVector1       60.       0.      0.
cellBasisVector2        0.      60.      0.
cellBasisVector3        0.       0      63.
cellOrigin               16.0       11.0      -29.1
}
wrapWater                  on
wrapAll                    on
# Force-Field Parameters
exclude                    scaled1-4
1-4scaling                 1.0
cutoff                     25.
switching                  on
switchdist                 20.
pairlistdist               30
# Integrator Parameters
timestep                   2.0    ;# 2fs/step
rigidBonds                 all    ;# needed for 2fs steps
nonbondedFreq              1
fullElectFrequency    2
stepspercycle              10
```

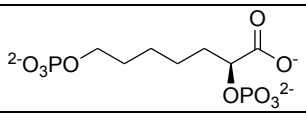
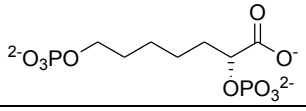
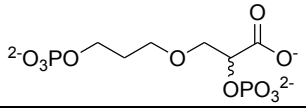
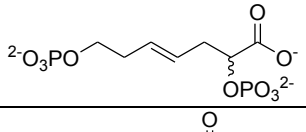
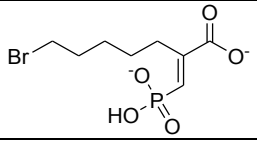
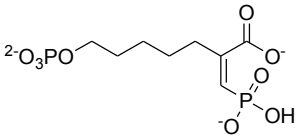
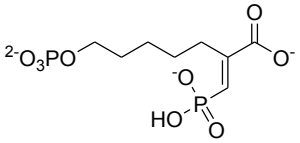
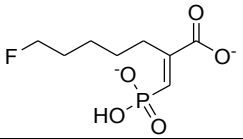
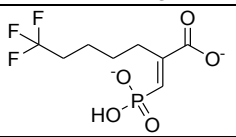
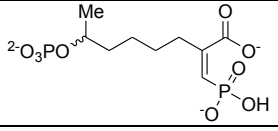
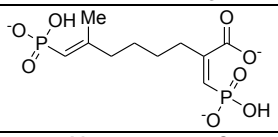
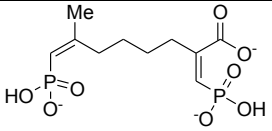
```
#PME (for full-system periodic electrostatics)
if {1} {
PME                      yes
PMEGridSizeX             40
PMEGridSizeY             40
PMEGridSizeZ             42
}
# Constant Temperature Control
langevin                 on       ;# do langevin dynamics
langevinDamping          5        ;# damping coefficient (gamma) of 5/ps
langevinTemp             $temperature
langevinHydrogen         no       ;# don't couple langevin bath to hydrogens
# Constant Pressure Control (variable volume)
if {1} {
useGroupPressure         yes ;# needed for 2fs steps
useFlexibleCell          no    ;# no for water box, yes for membrane
useConstantArea          no    ;# no for water box, yes for membrane
langevinPiston           on
langevinPistonTarget     1.01325 ;#    in bar -> 1 atm
langevinPistonPeriod     100.
langevinPistonDecay      50.
langevinPistonTemp       $temperature
}
restartfreq              100       ;# 500steps = every 1ps
dcdfreq                  2500
xstFreq                  2500
outputEnergies           100
outputPressure           100
#####################################################################
## EXECUTION SCRIPT##
#####################################################################
# Minimization
if {1} {
minimize                 5000
reinitvels               $temperature
}
run 2500000 ;# 5ns
```

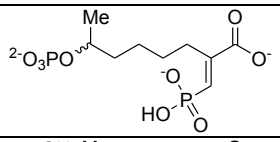
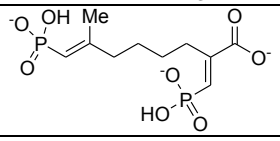
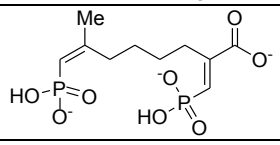
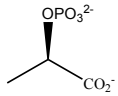
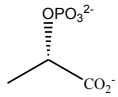
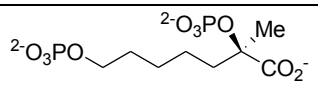
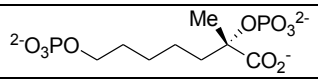
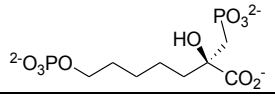
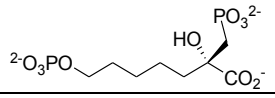
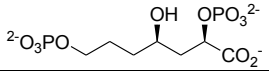
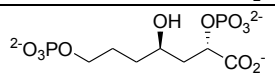
Appendix 2: Number of hydrogen bonds and van der Waals contacts from modelling of inhibitors into intermediate-adapted receptor grid ("good", "bad" and "ugly" van der Waals contacts are defined with cut off ratio of 1.3, 0.89 and 0.75 respectively)

Number	Structure	Hydrogen bonds	Good vdW	Bad vdW	Ugly vdW
S-5.1		14	284	18	11
R-5.1		15	291	27	11
S-5.3 *R*-5.3		14 14	291 283	17 20	10 11
S-5.4 *R*-5.4		16 15	292 283	18 19	8 12
5.5		11	291	18	7
5.6		13	287	22	4
5.7		16	316	26	11
5.11		10	279	16	6
5.12		11	265	21	5
S-5.13 *R*-5.13		12 11	313 324	17 23	8 8
5.14		11	289	20	9
5.15		10	311	14	5

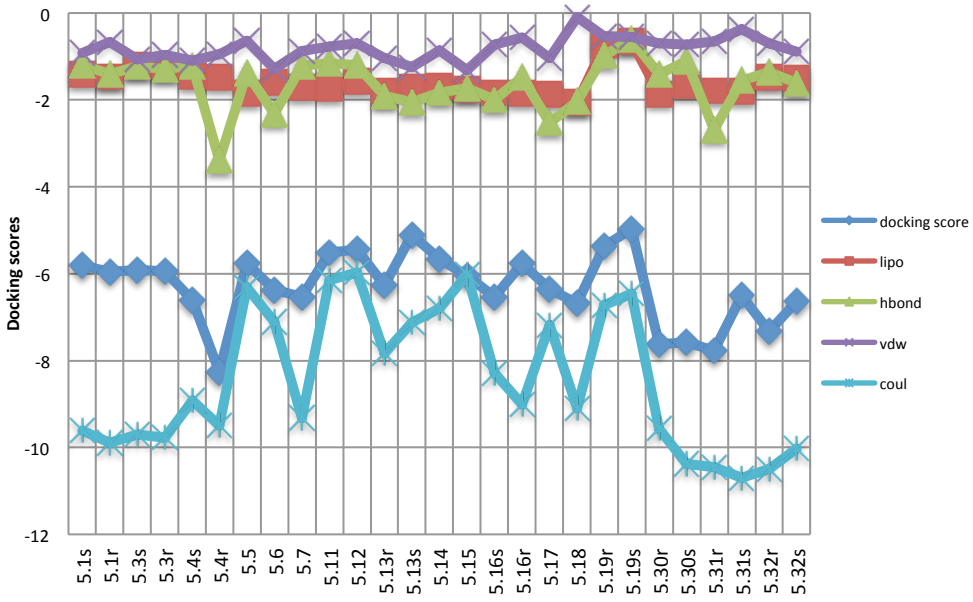
S-5.16 *R*-5.16		17 16	312 302	30 33	10 11
5.17		17	310	29	9
5.18		14	334	26	11
R-5.19		11	184	17	7
S-5.19		10	164	11	7
22s		17	305	25	10
R-5.30		14	352	28	12
S-5.30		16	319	24	13
R-5.31		19	321	31	13
S-5.31		15	326	32	12
R-5.32		16	309	20	13
S-5.32		16	318	26	10

Appendix 3: Number of hydrogen bonds and van der Waals contacts from modelling of inhibitors into wild type receptor grid based on crystal structure 2B7O ("good", "bad" and "ugly" van der Waals contacts are defined with cut off ratio of 1.3, 0.89 and 0.75 respectively)

Compound number	Compound structure	H-bonds	Good vdW	Bad vdW	Ugly vdW
S-5.1		8	231	13	6
R-5.1		10	230	11	5
S-5.3 *R*-5.3		8 8	230 230	13 11	6 4
S-5.4 *R*-5.4		8 9	240 233	12 11	5 4
5.5		6	239	10	5
5.6		7	222	10	3
5.7		9	253	14	7
5.11		6	214	7	2
5.12		7	216	8	4
S-5.13 *R*-5.13		7 7	281 248	8 10	6 4
5.14		9	248	15	6
5.15		8	224	12	6

S-5.16 *R*-5.16		9 8	263 282	11 12	5 4
5.17		9	257	13	7
5.18		8	286	9	6
R-5.19		8	154	7	4
S-5.19		7	145	7	6
R-5.30		7	268	15	5
S-5.30		8	259	11	7
R-5.31		12	250	21	6
S-5.31		9	231	20	5
R-5.32		11	251	14	5
S-5.32		10	254	11	7

Appendix 4: Plot of components of the docking scores for the best poses of intermediate mimicking structures



Note: "docking score" is the total score obtained from docking of structures into intermediate-adapted receptor; "lipo" is the score from lipophilic interactions; "hbond" is score from hydrogen bonding interactions; "vdw" indicates score from van der Waals interactions; and "coul" indicates score from electrostatic interactions.

Appendix 5: Amino acid sequences used in SCA

CME2
MLPGAAASEAVAGSVASKRPSSAITCTIRSMEGFVVSNPFRRRVSGRAAWQAV
RRCGSVALSSRVYYYVERGAEVCVRLGDGCGAFGLRNSVSIGGRQVGSEPWSPS
SWRQRPAHQLPVYPDKDQLALVEAGLRTLPGLVTPKSIRSLHEHLAAACHGEA
FVIQGGDCAETFSSRTLEYNISLYHLLAAQAAILSSTGSKRRIVRIGRIAGQFAKP
RSAPVDADGLPSFRGDIVHSAAPNGRARTPDPQRLLQAYRHASASLEHLRVLEH
LHASPMESLRILSEKLAWDSVVDSNLVKAIQNLVQGIDEEIIVMSSTGRHVTSID
TNVIENVFTSHEALLLPYEEGLVRLADWSPSGNAHTLEHLNRDAGSPIDLLSAPQ
YYATSAHFLWIGERTRQIDGAHVEFLRGVSNPIGVKLSGNATLEQVFALAEILNP
ANTPGRLTLITRMGVDALQRKLPPLMAAVKRSGLAVVWMCDPMHANTQSVR
DKNGNSFKTRDCEQVLDEMRLFFEITAAEGVPAGGIHLEMSGEAVTECLGGAT
RTAVHDLARAYHSACDPRLNYAQALMLTLALKDSLQWREALPNASSAVLSA

CRE1
MKLQTSQGQALRTSRPAVRGLPALSRRPVRVSAIAAPEKQSTPASLNGTSNWQP
TSWRSKPVVQQPEYLDKQEVVKACEEIARMPPLIFAGECRTLQSRLAKASTGDA
FVLFGGDCAEAFSQFSANRIRDLYRLLLQMSIVLAFGGGVPIVKLGRIAGQFAKP
RSASTETIGGVTLPAYRGDIINGPEFTEEARRCDPWRLVRAYNQSAATLNLLRGF
SYGGYAGLSRVSQWELEFMKNTPEGHAYMDIAKRVDEAIQFMIACGMDTNMP
FMRETEFYTCHECLLLDYEEALTRLDSTTDKWYGCSAHFLWCGERTRQLDHAH
VEFLRGVNNPIGVKVSDKMDPNDIVSLIASVNPSNTPGRLSIIVRMGAKALRAKL
PALIEAVQRSGQVVTWVCDPMHGNTETVSGFKTRRYENIRSEIEAFFDVHEKM
GTVPGGVHLEMTGDNVTECIGGGASISEDDLNSRYHTHCDPRLNAEQSLEIAFY
VAQRLRQRRENLAAKTA

OLU1
MGRGRTTTNARTNPERGARDASRAQATASKSKARSADVSATGANLKEWNPKS
WRQREALQQPNYENQAELEEALKVIANRPPLVFAGEARDLQEKLANAAAGNA
FVLFGGDCAESFRDFTSDNVRDTYRVLLQMSVVLMYGSGVPVVKLGRMAGQF
AKPRSEDLETIDGLSLPSYRGDNINSCEFTPEARRPDPSRLVKAYDQSCATLNLL
RAFSNGGYAAMTRVSDWNLDFMENTERGSQYEDLAQRVDAAIDFMAACGIDE
THPSMQETSFFTAHEALHLGYEESLTRLDSTTEEHYGCSAHFLWCGERTRQPEG
AHMEYFRGISNPIGIKISDKSDGEGVVSLVKKLNPDNVPGRITLISRMGAAKLRE
HLPRLITAIEDAGLNVLWVTDPMHGNTIKTDNGFKTRPFEAVRDEIMAFFEVHE
KMGTYPGGVHLEMTGQNVTECTGGIMDVSVSDLEKRYLTHCDPRLNASQAIEL
AFLMASELNDMRRRRAAQ

PTI1
MMLKLASALLLLASSEAFTPQSLVGRRQSALSVATEPPASTAATTGGNTDWSPR
SWREREVQQGPNYEDEEELEQAIDTIKKFSPLVFAGEVRSLHEQLARACSGQGF
LLMGGDCAEAFNEFNVDHVRDSFRVLLQMALVLTFGSAMPVIKVGRMAGQFA
KPRSEPDEVRDGVALPSYRGDIINREEFTPEARRHNPNNMVEAYHQSAQTLNIL
RAFSTGGYADMSRLHAWNLDFVETTDEGSRYRKFATKVDESLRFMKAIGVDTS
SPTFTKTEFYTAHECLLLPYEEALTRKDSTTGRYYDCSGHMLWVGERTRQLDG
AHLEFVRGIGNPLGVKISDKCTPEELIRIIDTMNPQNIPGRLTIVVRMGAEKVRKN
LPALIRAVQREGKSVLWISDPVHGNTYKTDSGIKTRNFDAIRDELRAFFDVHDE
MGSHPGGVHLEMTGEDVTECTGGISGVSEDTLNDRYHTFCDPRLNGAQALELA
FLIAERMRLRTGLPPIE

TPS1
MVSSAPKVANGSTSSGEGGEWSPSSWKNAVAKQMPIYEDEAELEDAIKALDRV
APLVFAGEVRSLHEKLARVSQGQGFVLMGGDCAESFKEFHVNHIRDTFRVIMQ
MALVMTFGSSMPVVKIGRMAGQFAKPRSEPNEVIDGVSLPSYRGDNVNREEFT
PEARRNDPHLMVDAYYQSAQTLNILRAFSTGGFADISRLHAWNLDFVENTEEG

SRYRLLAAKVDESLRFMKAIGIDTSKPEFTSVDFYTAHECLLLPYEQALTRQDST
TGKWYDCSSHMLWVGERTRDLDGAHLEFTRGIGNPLGVKVSDKCTPEELIRIIE
TMNPQNVPGKLSIIVRMGAEKLRKNLPGLIRAVQREGKAVVWISDPVHGNTRK
TESGFKTRDFDNIRAELRAFFDVHDEMGSHPGGVHIEMTGEDVTECTGGVSSVT
DETLKTRYNTACDPRLNGSQALELAFLIAERMRLRTGLPPLTSREAVSWRRLLG
IY

AMC1
MSNWQVDSWRNKPILQQPEYEDKVRLKEVEQTLSSYPPLVFAAEARELRRQLG
EVSLGKGFLLQGGDCAESFDEFNAPKIRDTFKVILQMAIVLTFAGRCPVTKVAR
MAGQYAKPRSSDFETRGGVKLPSYRGDIINNFEFTEAARRPDPERLLEAYHRSA
STLNLLRAFAQGGLADLHEVNRWNMAFVENNPLKDRYSDMANRIQDTLEFMD
VIGINAQTSSTLHETSLFTSHEALLLNYEQALTRVDTLTGKPYDCSAHMVWIGE
RTRQLDHAHVEFFRGIHNPIGVKVGPSMQEDELIRLIDALNPLNDPGRLTLITRM
GADNLADNLPRLLRKVKAERRHVVWSSDPMHGNTFSASSGYKTRNFDSILREI
KQFFQAHEAEGTHAGGIHLEMTGQHVTECTGGAYQISDDDLAQAYKTQCDPRL
NADQVLEMAFLVADHLRNA

GEM1
MPSQKWTKSSWRSFRALQQPVWPAGSALEETQKTLSQLPPLVFAGECQTLKAQ
LADAVEGRAFVLQCGDCAEDFSRCTGPDIRELLKVILQMSVVLAFAGEKRVIKI
GRIAGQYAKPRSSDTEMVHGIELPSYRGDMVNSPEPTLEARTPDPRRMLEGYYR
AAATLNLVRSFTLGGYASLERVQAWHRASLDALPAGQKYEDLVRQIWKTINF
MTAIGLDPQHTPQLNQVTLYTSHEALLLDYEEALTRMDSTSGGWYDCSAHML
WIGDRTRQLDGAHVEFLRGVKNPLGMKVGPSYDIDTVKALAQRLNPDNEPGRL
TLITRFGADKIDSYLPRLLKEMKQEGFKVVWSCDPMHGNTYQNEYGQKSRKFE
DILREIKNFQQIHKAEGTVAGGVHLELTGDHVTECTGGSRQLLDKHLHLNYQT
NCDPRLNAEQSVELAFELAEMLHPCK

DPS1
MSQSQKEWNKGSWKNYTALQQPNWPNEGAVEKVLANLSQLPPLVFAGEIRNL
KQQLAKAVTGEAFLLQGGDCSENFSQVTAPKIRETLRVLLQMAVVLTYAGDKP
VIKVGRLAGQFAKPRSSDTETINGVTLPSYRGDMVNCPDFTETCRVPNPKLMLK
GYNMAASTLNLVRAFTRGGFGSLHRVQAWNQEFVAQSPMGRSYDRLAKQIDQ
TIKFMGNIGIAADNPLIEQTQLFTSHEALLLQYEEAMTRVDSTAGGWYDCSAHM
LWIGERTRQLGGAHVEFLRGVLNPIGIKVGPNNDPDDIKRLCETLNPENEAGRIT
LITRLGVKEIEKSLPPLLREMKREGHNIVWSCDPMHANTYTATSGHKTRNFNDI
LSELTCFFESHWSEGTIPGGVHFEMTGEDVTECTGGARNIVDEELARNYLTTCD
PRLNAEQSLEVAFQIADMIRS

DAT1
MVNTKEWTKSSWKNHTALQQPNWPDKKKLDKVIDDLTLMPPLVFAGEIRALK
EMLGKASKGEAFLLQGGDCSEDFSQVTAPNIRESLKVLLQMAVVMTYAGGKPT
IKVGRIAGQFAKPRSSDTEMVNGVELPSYRGDMVNKVTPTIKDRTPNPKFMLK
GYYMAAATLNLLRAFTRGGYAALHRVQSWNHEFVVQSPMGRSYDRLARQIDK
AIKFMNTIGIPTDIPQVNQTQFFTSHEALLLPYEEALTRIDSTTGGWYDCSAHML
WIGERTRQVDGAHVAFLKGVLNPIGVKIGPNYDMDNVKQLIETLNPENEAGRL
TLITRMGCNNIEKKLAPLLRETKKQGYHIVWNCDPMHANTYTSESGHKTRDFN
DILKEITRFFEIHWAEGTIPGGVHLEMTGKNVTECVGGARNIVSEELHNRYDTTC
DPRLNAEQSLEVAFQIADMIKH

MGM1
MSNPWQPDSWRKFNALQQPEWPNQDELQQVCKNLSSYPPLVFAGEVRSLSGH
LKRVANGKAFLLQGGDCAESFGEFNANAIRDKLKILLQMAVILTYGGGRPIVKV
GRIAGQFAKPRSSPTETQGDQTLPSFRGEMVNDPDFSESSRNPEPNRLERVYFQS
ASTLNLLRAFTSGGFADLHSLNNWTRDFVSNSPQGKRYAEIADKLTDALKFMD
TVGINSTNTPVLHEVEYFTSHEALILDYEQALTREDSLTNEYYCCSAHMLWIGE

RTRQLDGAHVEFLRGVKNPIGVKLGPSATAEDALRLCDALNPQNIPGRLTFITRF
GHNKVEENLPKLIRSIKQAGRSVIWSCDPMHGNTFTASSGYKTRNVDHVMSEIR
SFFAVHKAEGTCPGGVHFELTGDAVTECVGGSHQVTEAHLPERYETTCDPRLN
ATQSLDIAFLITETLQDFKK

GOX1
MHPSSQTPTSTNTMSWSPASWRSRPIVQAPSYADEAALAAVEARLHRYPPLVFA
GEARRLKTRLAAASRGEAFVLQGGACAESFDEFTADIVRDTFRVLLQMAVVLT
FAAKVPVVKIGRMAGQYAKPRSSNTETVDGVTLPCYRGDIINGPEFTPEARIPDP
ARMETGYFQSAGVMNLLRAFAHGGYANLHQVHRWNLGFVERSPLAARYQDL
AHRIDETLSFLRACGFDGSASQIDETEFYTSHEALLLPYEQALTRVDSTSGDYYD
CSAHFLWIGDRTRQPDGAHVEFLRGVQNPIGIKVGPTTTVADLEKLLEILNPKDE
AGRITLISRMGKDKVRDHLPPLLDAVNRTGRTVTWLCDPMHGNTTTTASKIKT
RSFDNILGEVLGFFDVFEAIGRRPGGIHLEMTGQDVTECIGGAQCLTEADLGGR
YETFCDPRLNAEQSLEMAFLLAQELTGRAGKKE

GBE1
MCRVAWKDCPAGAIEKCRLRLDRTEFAGTRYAMDTIMTAGPSAPGAAKAWTP
QTWRERPILQVPVYPDPAALSEVEQRLHRYPPLVFAGEARRLKAQLAEAAQGR
AFVLQGGDCAESFGDFTANIIRDTFRVLLQMAVVLTFGAGLPVVKVGRMAGQF
AKPRSSDTETVNGETLPSYRGDIINGPAFTAEDRVPDPARMETAYFQSAGRMNL
LRAFASGGYADLHEVHRWNLGFVERSPLASQYRDMASRIDETLRFMGACGLK
DTAQVRETEFYTGHEALLLPYEEAFTRIDSTTGEYYGCSAHFIWIGERTRQPDGA
HVEFMRGIHNPIGIKLGPNADPDEVVRLCELLNPRAEPGRITLISRMGADKVRDR
LPPLLRAVKRSGVPVVWVCDPMHGNTISTTSKVKTRNFDSILGEVRGFFDAHAQ
EGTWAGGIHVEMTGRDVTECIGGAHRLTETDLGRAYETFCDPRLNAEQSLELA
FLLASELKQRLHHVSKSIP

ACR1
MNDADTTMHDPAGWSPESWRAKPIKQVPTYPDAAKLAAMEARLHGYPPLVFA
GEARRLKASLALATEGKAFILQGGDCAESFGDFTANVIRDTFRVLLQMAVVLTF
GGSVPVVKLGRMAGQFAKPRSSDSETIDGVTLPSYRGDIINGPDFTAEARIPDPA
RMEFGYFQSAGTLNLLRAFASGGYADLHQVHRWNLDFVRRSPLAERYRDLAA
RIDETLQFMAACGMTSQSTPQIHETEFYTSHEALLLPYEQALTRVDSTTGDHYA
CSAHFLWIGDRTRQLDGAHVEFMRGVKNPIGMKVGPSMDEDDMLRLIERLDPA
NEAGRLTLIARMGAEQIAAKLPKLVRATKREGRNVAWLCDPMHGNTITTSGKV
KTRRFDAILSEVRGFFDIHEAEGTTPSGVHVEMTGQDVTECVGGARQLSEADLG
ARYETFCDPRLNAEQSLELAFLIAAELKTRRAPRGTAIAAE

MAG1
MSENWSPESWRAKPIHQAPTYPDSAKLAEVEESLRNFPPLVFAGEARRLKASLA
DVADGKAFLLQGGDCAESFAEFHANNIRDTFRVLLQMAVVLTYGAAMPVVKV
GRMAGQFAKPRSADTETIDGVTLPSYRGDNVNGSEFTAEARVPDPLRMIRAYN
QSAATLNLLRAFAQGGYADLHKVHQWTLGCVAGTLQAARYQDLCDRLDETL
AFMEACGMTSDTTPQLRETDFFTSHEALLMPYEQALTRIDSTTGDWYDCSAHM
LWIGERTRQMDAGHVEFLRGVKNPLGFKAGPGMSPDDLLRLIDRLNPENESGRI
TVITRMGAEKVEQVLPGLIRAVEREGRKVVWSCDPMHANTIKAAGGYKTRPFD
AILAEVRAFFAVHKAEGSHAGGVHFEMTGQDVTECIGGMTGVAEEDLGDRYQ
TACDPRLNATQSLELAFLIAEQLKDERALLRAKAKAKRKD

RCE1
MTGLAASSGARSGKRCEMATETAMGTGGRWTPESWRAKPAMQMPTYPDPAK
LQAAAARLAGYPPLVFAGEARSLKECLGRATQGQAFLLQGGDCAESFSEFHPN
NIRDTFRVLLQMAIVLTFGGATPVVKVGRMAGQFAKPRSADMEVVDGVELPA
YRGDIINGFDFTPEARVPDPDRMVQAYNQSAATLNLLRAFAQGGYADLHKVHQ
WTLGFVEKSPQGERYRDLAGRLDEALAFMAACGITGATTPQIRETEFFTSHEAL
LLPYEQALTRIDSTTGDWYAVSAHMLWVGDRTRQPDGAHIEYLRGIRNPLGLK

CGPSLDPDELIRLIDLLNPADEPGRLTLICRMGHDKVEDKLPALVRKVKREGRT
VVWSCDPMHGNTIKSSTGFKTRPFDRVLAEVRGFFEVCRTEDVHPGGVHFEMT
GQDVTECTGGAQAITEHSLALRYHTACDPRLNASQSLELAFLVAEELKRTRRGL
KLSQQAAAE

ELI1
MARNWTPDSWKNFEARHLPTYEDAAALRDAEATLAAYPPLVFAGEARALKAD
LAEVANGEAFLLQGGDCAESFAEFHPNNIRDTFRVILQMAVVLTFAGKLPVVK
VGRMAGQFAKPRSSDTETQGDMTLPSYFGDNVNGIDFDPAQRRNDPERMVKA
YSQAAATLNLLRAFAGGGYANLRQVHQWTLDFMGRSPWADRFGEMADRITE
ALDFMEACGVDPATVPQLQGTSFYTSHEGLLLPYEQAMARQDSLTGDWYDTS
AHMLWIGDRTRFEGSAHVEFMRGIGNPLGVKCGPSLEPDVLLKLLDDLNPARE
AGRITLISRFGHDKVEAGLPRLVRAVTREGHPVVWSCDPMHGNVIKSDSGYKT
RPFDRIKTEVRGFFDVHRAEGTHAGGIHLEMTGQDVTECVGGAVAITDEALGD
RYHTHCDPRLNAAQSIELAFAIADMLQLAAKEQQQADAA

SWI1
MATNWTPASWRQQQAVQMPDYPDAAALAAAEQELSSFPPLVFAGEARELTAS
LAQVAAGKAFLLQGGDCAESFAEFHPNNIRDTFRVLLQMAVVLTFASKLPVVK
VGRMAGQFAKPRSAPTEEIGGVTLPSYRGDIINDIGFSEESRIPDPRRMVRAYSQS
AATLNLLRAFAQGGYANLHQVHAWTLDFMGRSPWAAQYGEVAARIGDALDF
MAACGINPETVPQLKGTQFYTSHEALLLPYEQAMTREDSLNTGQYFDTSAHFL
WIGDRTRFEGSAHVEYLRGVGNPIGLKCGPSLEPDALLRLLDTLDPQRQPGRITL
ITRYGHDKIEKHLPALVRAVKREGRPVVWSCDPMHGNVVKAANGYKTRPFDRI
LAEVRGFFAVHRAEGTHGGGIHAEMTGQNVTECTGGAMAVTDENLADRYHTH
CDPRLNAGQSLELAFLLAEMLNQEMKERDRAAA

PLA1
MADTWKPESWRAKPAKHLPSYPDEAALAAVEARLRSYPPLVFAGEARKLKAD
LAEVCEGRAFLLQGGDCAESFAEFSADNIRDTFRVLLQMAVVLTFAAASPVVK
VGRIAGQFAKPRSSATETIGDVTLPSYLGDNINGIEFDEKSRVPDPERLLRAYSQS
ASTLNLIRAFANGGYADLDFVHRWNLGFVSDSAEGARYEELANRITEALDFMR
ACGIDSATQPQLHTTDFYTSHEALLLGYEQAMTRIDSTTGDWYDTSAHMLWIG
DRTRQPDHGHVEYMRGIKNPIGMKCGPSLDPEELVRLTDILNPKNEPGRLTLICR
FGAENVEKHLPQLIRAIEREGKKVVWSCDPMHGNTIKASSGYKTRPVDRILAEV
QAFMAVHRAEGTHAGGVHFEMTGQNVTECIGGAQAISETQLGDRYHTHCDPR
LNASQSLELAFLIAEGLKKERLEALRAEPVAALGAW

SIT1
MSDWQKTNWRSKPRVQMPDYTDQAALQAVEAQLAKYPPLVFAGESRRLKAQ
LGAAGRGEAFLLQGGDCAESFEQFSADGIRDTFKVMLQMAMVLTYGAKVPVV
KVGRMAGQFAKPRSAPTETVDGVELPSYRGDIINELAFTPEARIPDPRKMLQAY
TQAAATLNLIRAFSTGGYADVHQVHAWTLGFTEGDKAEAYRDMANRITDTLD
FMKAAGVTADNAHTLQTVEFYTSHEGLLLEYEEALTRLDSTSGKWLAGSGHMI
WIGDRTRQPDGAHVEFCSGVLNPIGLKCGPTTTADDLKVLMQKLNPENEEGKL
TLIARFGAGKVADHLPRLIQAVKDEGANVTWVCDPMHGNTIKSASGYKTRPFD
SVLREVRDFFGVHQAEGTIPGGVHFEMTGQDVTECTGGVREVTDEDLSDRYHT
ACDPRLNADQSLELAFLVAEELSRLRTPDDTRAAI

DSH1
MTKAWSKSDWRTKPRIQMPEYMDPAALAAVEARLTQYPPLVFAGEARSLRRE
LADVANGKGFLLQGGDCAESFGEFGADLIRDTFKVMLQMAMVLTYGAKVPVV
KVGRMAGQFAKPRSAPTEVKEGVELPSYRGDIINDLDFTPESRIPNPEKMLQAY
TQAAATLNLLRAFSKGGYADIHQVHAWTLGFTDRDEAEKYREMATRIADALD
FMKSAGLTSENNSELATVDFYTSHEALLLEYEEALCRIDTTTGLPLAGSGHMLW
IGDRTRQPDGAHVTFCAGVQNPIGLKCGPSITTDDLKVLMARLNPKNEAGRLTL
IARFGAGSVGDHLPRLVKAVQEEGANVVWSCDPMHGNTIKSASGYKTRPFESV

LREVQEFFAVHNAEGTYPGGVHFEMTGKDVTECTGGVRAVSDEDLSSRYHTA
CDPRLNASQALELAFLVAEEIETHRGATGAAKRTA

PDE1
MAGTDCMKDRIMAPTNSSSATPAQAWDKRGWRAYPRVQMPDYPDQPGVAAV
EAQLAKYPPLVFAGEARRLKASLGEVAAGRAFLLQGGDCAESFAEFSADNIRDT
FKVMLQMAVVLTWGAQLPVVKVGRMAGQFAKPRSAPTEKMGEQELPSYRGD
IINGFDFTPEARIPDPQRMLAAYTQAAASLNLLRAFSTGGYADIHRVQSWISDFT
GGEEAARYRDIAERISDAMAFMAAAGVTSETAHDLGKVDFYTSHEALLLEYEE
ALARIDSTTGLPVAGSGHMIWIGDRTRQVDGAHVEFCRGVQNPIGLKCGPSISD
SDLKALIAKLNPENEPGRLTLIARFGAGSVGDHLPRLIRAVREEGANVVWSCDP
MHGNTIKSASGYKTRPFDSVLREVREFFAVHRAEGTIPGGVHFEMTGQDVTECI
GGVRAVTDEDLSNRYHTACDPRLNASQSLELAFLVAEELRNRRVNGASKQAK
AG

MXA1
MTNPTWTPDSWRHKPVKYMPEDYPDAQALVQTEAELSRLPPLVFADETRQLT
ARLARVAEGKAILLQGGDCAESFKEFTTDNIRDTYRLLLQMAMVLTFAGNRPV
VKVGRIAGQFAKPRSSPMETLGGVTLPSYRGDIINGMEFDAQARTPDPRRLLKA
YHQSSATLNLLRAFAQRGYEELAEPQRWPQGGINRLLADRILESLSFMRALGVS
PEPRSGSNAIDFFTSHEALLLNVEEAMTRAEPDGSGWYDSSAHMLWIGERTRQL
DGGHVEFMRGIQNPIGIKCGPTMEPDEVLRLMDLLNPQGIPGRITLIGRFGADLV
AERLPRLMAATRRDGRPVIWSIDPMHGNTHKAGNGYKTRSFDRILTEVCGFMD
AAAAEGVHPGGLHLEMTGQNVTECLGGPQPVTEDDLSSRYHTHCDPRLNADQ
ALQLAFLVAESMPSLRTHEARAA

HNE1
MTWHPSDWRNLPAKHIPQDYPDAAALAAVETRLRSYPPLVFAGEARRLKTRLA
EVAAGKAFLLQGGDCAESFKEFHPDNIRDTFRVILQMAVVLTFAAAKPVVKVG
RIAGQFGKPRSEPIETIDGVTLPSYRGDNINGMEFTPESRIPDPERLIQAYSQSAAT
LNLLRAFSQGGYASLANVQRWMLGFVDRSPQGERYQKLADQISASLRFMEAV
GLNAESVPQMSQVEFYTSHEALLLGYEEAMTRIDSTSGDWYDTSAHMLWIGHR
TRQIDHAHVNFCKGVKNPIGIKCGPGMETDELLRLIDTLNPADEPGRITLISRFGS
EGVKAGLPPLARAVKKSGRTVVWSCDPMHGNTLKTGTGFKTRPVDRILSEVRQ
FIDVLSAEGCYPGGVHFEMTGQNVTECIGGAQAISEEDLSSRYHTHCDPRLNGE
QALELAFLVAEKLQAAEVEAATARKAG

HBA1
MTKWAPDSWRAKPAKHIPEDYPNSAALADVEATLSQYPPLVFAGEVRRLKARL
GDVAAGRAFLLQGGDCAESFKEFNSNNIRDTFRVILQMAVTLTFAASKPVVKV
GRIAGQFAKPRSSPVEVKDGVTLPSYRGDIINGMEFTPESRTPDPQRLIRAYNQS
AATMNLVRAFSRGGYADLHNLNKWMLSFVDRSPQGERFQKIATQINEAVDFM
EAIGLTPDNTPIMSETEFYTSHEALLLGYEQAMTREDSTIPNQTVDTSAHMLWIG
DRTRQLDGAHVEFCRGVINPIGMKCGPTLDPDELIRLIDTLNPEDEPGKMVLIAR
FGADKVEDGLSKLVRRVEQEGRSVIWSCDPMHGNTLTSESGYKTRPVDRILSEV
AQFIDVVSTEGAYPGGVHFEMTGQNVTECLGGAQDITDADLSSRYHTHCDPRL
NGEQALELAFLVAEKLRSKKSSVALAAQ

MMR1
MTWSPQSWRSKPVSQMPNYKDAAKLDAAIAELSARPALVFAGEARRLRRQLA
DVTAGKAFLLQGGDCAESFKEFSTEGVRDTFRVLLQMAVVMTFAASKPIVKVG
RIAGQFAKPRSADMETIDGVSLPSYRGDSVNGPEFTPEAREPDPQRLIRAYDQSA
STLNLLRAFASGGYADLHNVHQWTQDFVSDSPAAERYAETAARISEALAFMKA
CGIGRDSAPSLEAVDFFTSHEALHLPFEEALTRRDPNTGQWYATSAHMIWTGER
TRQLDGAHVEYARGIANPVGVKCGPTMQPDDLLPLIDALNPDNEAGRLVLIVR
MGADNVVKNLPKLAAAVTKAGRKVVWSSDPMHGNTHKTSNGYKTRDFDRIL
SELEGFMDVLYAEGAYPGGVHFEMTGRDVTECVGGAKTVTEADLAARYHTHC

DPRLNADQALDMAFRIAESLKRVRNNNSAANAA

PIN1
MTQWTPSSWRNKPVQQLPSYPDQDKLQQTEHYLANQPPLVFAGEARALREDL
AKVAKGEAFLLQGGDCAESFDEFRTSHIRDTFKALMQMAVVLTFGGQKPVVKI
GRIAGQFAKPRSADMETINGLSLPSYRGDIINGIDFNEKSRLPDPQRMVTAYNQS
TSTLNLLRAFAQGGLADLNQVHQWNLDFIKKSPLGKRYETLAGSIEDSLAFMN
ACGINPATAPQLRETTLYTSHEALLLPYEQALTRCDSLTGKWYDCSAHMLWIG
DRTRQIDHAHIEFLRGVENPIGIKVGPSTVPEELLRIIDIVNPNNEAGRLNLIVRM
GADKVADHLPALIGAIEREGKNVVWSSDPMHGNTIKAPNGYKTRRVNDILREV
QQFFQIHKAEGSYAGGVHFEMTGRNVTECVGGAFQITEHDLAQRYHTYCDPRL
NADQALELAFLIAEHIKAAR

MMW1
MSQWNLTSWREKTALQQPVYPNAEHLAQVENTLGKMPPLVFAGEARQLKKAL
AQVANRQSFLLQGGDCAESFAEFHANNIRDTFKVMLQMAVVLTYAGKCPVVK
VGRMAGQFAKPRSSGSEVIGGIELPSYRGDIINGIDFTEQARVPDPERLVQVYNQ
SASTMNLLRAFAQGGFADLHQVHQWNLDFLNASPAGSRFQGVADKIDDALQF
MEACGIGPGLAQLKETDFYTSHEALLLPYEQALTRKDSLTGDWYDCSAHMLWI
GDRTRQLDGAHVEFLRGVQNPIGVKAGPTMDPEDLLRLCDVLNPNNEAGRLNII
VRMGADKVEDGMPKLIQAIQREGKQVVWSSDPMHGNTVKASTGYKTRRVDD
VLKEVQQFFQVHNAEGSYAGGVHFEMTGQNVTECVGGAFEVTEADLADRYHT
HCDPRLNADQSLELAFMISETLKKARS

TDN1
MSNWSASSWRQKPIVQQPTYPNQDELNRVLEELRNYPPLVFAGEARSLKAQLA
NASEGRGFLLQGGDCAESFSEFHAHNIRDTFKAILQMAVVMTYAGGVPVVKVG
RLGGQFAKPRSSDTETIDGITLDAYRGDIINGVDFTKEMRVPDPQRMIKAYNQS
AATLNLLRAFASGGLADLHVVHKWNLDFTNQSEISKKYEDLAEEIGKSLKFME
ACGITSKTHRTLRETDFYTSHEALLLPYEEAFTRKDSITNEWYDTSAHMLWIGD
RTRQLDGAHVEYMRGINNPIGVKAGPSMDPEDLLRLCDVLNPQNEAGRLNIIVR
MGADKVADGMPKLIRAIEREGRKVLWSCDPMHGNTVKSSNNFKTRPVDSILTE
MKQFFQVHKSEGTFAGGVHLEMTGKNVTECIGGSFVVTEEDLSSRYHTHCDPR
LNADQALELAFLIADTLREAQK

WSU1
MQNEWNPQSWREKPIRQQPTYKDREHLKNIERELKQYPPLVFAGEARSLKKRL
ADVTEGRAFLLQGGDCAESFSEFNAINIRDMFKVILQMAVVLTFAGSCPIVKVG
RLAGQFAKPRSSDEEVMDGVSLPSYRGDIINSIEFTPEAREPNPKRMLKAYNQSA
ATLNLIRAFAQGGLADLHEVHRWNLGFIESSFGERYQALCDRITETLSFMEACGI
TSATTPMLHETEFYTSHEALLLNYEEALTRKDSLTGDWYDCSAHMLWIGERTR
ELEGAHMEFLRGVNNPVGVKVGPGATKEELLGICDILNPLNEAGRLNFIVRMG
AEKIGEKLPALLRIIKEEGRKVLWSIDPMHGNTVKASSGYKTRAFNQVLSEVRSF
FDIHKAEGTYAGGIHLEMTGQDVTECIGGSQKITEEGLACNYNTQCDPRLNATQ
ALELAFLIAEILKKR

CHA1
MQWNRESWQNFNILQQPIYENSDKLQSCKNQLKKLPPLVFAGEVRALKDELKS
ATLGKSFLLQGGDCAESFSNFSANNIRDMFKLMLQMAIILTFAGGVPVVKIGRIA
GQFAKPRSSDFEEIDGVELPSYRGDIINGFEFTKEARKPDPNRMIEAYNQSTSTLN
LLRAFSRGGLADLHEVHKWNLGFLKRGELAKKFQNLANEITRTLNFMEACGIN
SANTQNLRETVLYTSHEALLLHYEECLTRIDSLTGDPYDCSAHMLWIGERTRGI
NDAHVNFISGIKNPVGIKIGPNGTASEILALCDKINPNNEIGKMNIIIRMGADKIG
DRLPKILRELKKENRAILFSIDPMHGNTIKTANNYKTRKFNQILSEVKSFFEICEA
ESVYAGGIHLEMTGLDVTECTGGAFNVTEETLKKRYETQCDPRLNADQALELA
FLIADNLKKRVIKEQ

NIS1
MSWSPSSWRNFPIKQQPDYKDKELLRKIEEEIKTFPPLIFAGEARELKSKLAKVA
KGEAILLQGGDCAESFANFNAANIRELFKVLLQMNMVMMYATGKPVVKIGRIA
GQYAKPRSSDFEEINGIKLPSYRGDIVNDVEFTPEAREPKPQKLLEAYYKSAATL
NLIRAYARGGFADLRAIHKWNLEFLKGNELENKFKELADKITQALHFMEACGV
DPDNTPQLKQTTLYTSHEALLLGYEEALTRKDSFTGEWYDCSAHFLWIGDRTR
ALDGAHVEFFRGIKNPIGVKVGPSMQKDELLALIDKLNPENEPGRLTLIVRMGA
EKIADIFPPLLRAVKEAGKEVVWSIDPMHGNTYKTESGLKTRDFEKILSEVKQFI
QIHQSEGTIPGGMHLEMTGSDVTECTGSISASITEEGLQSRYHTQCDPRLNANQA
LELAFMVAETFKDLQK

SUN1
MSWTPGSWRDFPIKQQPTYQDQETLQKVEAELSSYPPLIFAGEARNLKRKLAAA
GRGEAFLLQGGDCAESFADFNAETIKNLFKLMLQMNMVLMYSTGKPVVKVGR
IAGQFAKPRSSDFEEVDGVKLPSYRGDIINGIEFTEEARIPNPHNMIRAYNQSAAT
LNLVRAFSRGGLADLNKVHQWNLDFIKDNPLGKRYDELSDKIDHAMKFMSAC
GLTSETMPQLHQTTLYTSHEALLLNYEEALTRLDTETGEWYDCSAHMLWIGDR
TRDLNEAHIEYFRGIKNPIGCKVGPSMGEDELIELIDALNPDNEEGRLNLIVRMG
AEKIAEFFPPLLKKVRDAGKNVVWTIDPMHGNVEKSSSGFKTRDFDNILSEVKQ
FFSIHKEMGTVAAGIHLEMTGNDVTECTGSTSCAITAEGLASRYHTQCDPRLNA
SQALELAFMLSDTISGADQ

ABU1
MNNWTPSSWRNFPIKQQPTYNDVETLAQVEKELASYPPLIFAGEALNLKKQLA
KVVNKEAFLLQGGDCAESFNVFNATNIKDLFKVMMQMAVVLTFSGGCPVVKV
GRVAGQFAKPRSADFEDINGLSLPSYRGDIINDIDFSLESREPKAKKLLKAYNQS
AATMNLLRAFARGGMADLNQVHLWNLDFVKDNTLGTKYEELANKISESLAFM
KACGITSENTPQLNQTTLFTSHEALLLNYEQALTRRDSITGDWFNCAAHMLWIG
DRTRELDGAHIEYFRGINNPIGCKVGPSMKEDELLKLIDTLNPNNEAGRLNLIVR
MGANKIADHFPKLLKRVEEEGKKVLWSSDPMHGNTIKAENGYKTRDFEAILSE
VKQFFQIHRAQGSYAGGIHLEMTGQNVTECTGSKSSAVTQGDLASRYHTQCDP
RLNADQALELAFMIADTLKEARK

NAM1
MKWTKESWRDKPIKQQPTYEDIEKLKQVEKELSKYPPLVFAGEVRNLKSKLEN
VCEGKAFLLQGGDCAESFAEFDGNNIRDMFKVMMQMAVVLTFAGGLPVVKIG
RIAGQFAKPRSSDFEERDGVKLPSYRGDIINSIEFTPEARKADPERMLKAYNQSA
ATMNLLRAFAKGGMADLRKIHKWTLDFVKESPEGHRFEELADRITESLKFMEA
VGITPENTPVLRETTLYTSHEALLLNYEEALTRQDSLTGDWYDCSAHMLWIGD
RTRDPNEAHVEFLRGVKNPIGIKAGPSMTPEGLIELIEKLNPENEAGRLNVIVRM
GADKIEKEFPKLLKAVVDTGKKVIWSIDPMHGNTIKTSSNIKTREFDKILAEVKG
YFAVHESFGTFPGGVHLEMTGRDVTECIGGSYSISEKDLEKSYETHCDPRLNAK
QALELSFLIADLLKQSINKNPQI

PUB1
MKNWKINSWRNYPVKHIPEYPDQKELDGVLSKIKDFPPLVFAGETRHLKEQLA
DVVDGKAFLLQGGDCAESFAEFHPDNIRDTFKLILQMSLVLTYSASLPVIKLGRI
AGQFSKPRSSPVENIDGVELPSYLGDNINGMEFNEKSRVPDPKRLFKAYSQSAST
LNLIRAFSHGGFADLKKVHTWNLGFIKNSPAAKKFKDLEDRIADALAFMDACGI
NSDFNRRLKTVNFWTSHEALLLPFEQAMTRVDSTTGEYHDTSAHFVWIGDRTR
QLDGGHVEFCRGIENPIGIKCGPTLKAEDLINLCNKINPNNEKGKITLISRFGHEN
VSKFLPKLIRAIKKEGLNVIWSCDPCHGNTIKATTGFKTRPFNSVVKEVKNVFEC
HQSEGSYAGGLHIEMTGQNVTECTGGAQKISDQDLSSRYHTHCDPRLNANQAL
ELAFLISDEIKKNAAYSKKNIKAAS

CME1
MAVCFVLPHNGLAAVAGRRAQSGWKPSGALFRETRMVISAKPDSKKNTGGRV

WRPSSWREMPVKQVPTYPDEARLQDVEARLHKALPLVYTGECEHLKKLIAKAS
RGEAFVLQGGDCAESFEEFEKYGGDNIRDTFLLLIQMSIVLMYALGLPVVKIGR
MAGQFAKPRTSPMEREKNKKLELPSYRGDMINGPEFEARARVPDPERMLRAYH
QSTATINLVRALAYGGFVDLHRIHEINLEFVNGTPQGKRFFEYAARITDALRFMS
SCGIQANSPMIKQAEFFVSHEALLLPYEEALVRQDRETGKYYATSGHMIWCGD
RTRQPDGGHVEFLRGIANPIGIKCGPSLTEAELIELLDILNPDNEPGRITLISRVGA
GRVREHLPRFIECVQREGRIVTWSCDPMHGNTVTTASGVKTRRFDDILSEVKEF
FEVHREMGSIPGGIHLEMTGQNVTECIGGMYELSDADLATRFESRCDPRLNASQ
SLELAFLLSEYMSGKRTD

ZMO1
MVTDWTPDSWRDFEARQLPIYPDNDKLDKALDQLSSYPPLILGSEARLLKEKLA
DVAKGKAFVVQGGDCAESFRDFSAESVRATLSALLWMDAVVSYASGKPVVRI
GRIAGQFAKPRSSNMEEKDGVSLPSYRGDNVNDISFTEEARLPDPDRLLAGHAQ
SAITLNLIRAYTQAGDDKWEYAFNGLHDFVNSRSWADDSWRGLVAEMGDAFR
FAQSCGGFGNGYAAPCFTSHEALHLPYEEAMVRSDCDYGLDYATSGHFLWIGD
RTRFEGSAHVEFLRGVDNPIGLKCGPNLEADELLRLLDVLDPSREAGRITLITRY
GADKIEKHLPTLVRAVKAEGRPVAWICDPMHGNGTTTSEGQKTRPFERILDEVK
GFFAVHQAEGTYAGGIHLEMTGLDVTECIGGSAGLSEEDLKRRYMTHCDPRLS
RDQALEMAFQVADLF

YLI1
MPPKVVITDPTPLKASEGAKSAAKLRSKSKEAAQLPKPNLGDAWAPTSWQSKPI
TQDVVYDDYTGVEAALEKLESLPPLVSPFEIEKLRTKLAHAANGKAFVLQGGD
CAELFDYCTQDRIESKLKVLMQMSLVLLWGTGLPVVRIGRIAGQYAKPRSKLTE
EVDGKTINSFRGDNINGFDVSDRKPDPARLVSSYFHSAATLNYIRCVLASGFADL
RKVDWSLEHVQNPETKERYQKTLDAITEGLQFMHTIGADTAQDMQTIDFFTSH
EALLLEYEQSLTREANGHHYNTSAHFIWIGDRTRQLDGAHVEFFRGVRNPIGIK
VGPTMKADELVRLLDIVDANHEDGRVTLITRYGASKVKELLPAHIKAVQSSGH
KVVWMSDPCHGNTQTSTVTKLKTRHVDDIVEELRQALEIHRECGSKLNGAHLE
LTGDAVTECIGGSAGLCDEDLVVRYDTVCDPRLSLSQSLDVSFLIADSMRAQK

CNB1
MAPAPAPWHPSSWREKPIAQDVVYEDKDQLETVLNKLRRLPPLVSPVEIDRLRT
QLADVAAGKAFLLQGGDCAELFDDCSQDPIEHKLSLILLMSLIILHGSRLPVVRI
ARIAGQYAKPRSKPTEIVEFPTKDGKTEKKEVLSFRGDNVNGYDPTDRAPDPQR
LLGSYFHSTATLNYIRTLLSSGFANLHNPVDWSFSHVRSPELQQAFSSVIESLQDS
LEFMKVATGAVGGGERGGMETVDFYTSHEALLLEYEEAFTRSWDSTTLSPPAT
GDSTPILSRSASRIREASSSSSYPHSPARRPKSPKNLSDSINSLSMSVGDLGKKEA
KKWYNTSAHFIWIGDRTRQLDGAHVEYFRGIANPIGIKIGPSMQPEEIVRVLDIV
NPDKIPGKVTLIGRYGAAKVDQFLPKHIDAVLRTDHPVVWQCDAMHGNTKSSV
HDPTLKTRHFVDVITEITRSMEIHKEKNTILGGVHLELTGEVNDDGYSVTECIGG
SMELEDKDLSFNYRTHCDPRLNYEQSLDVAFLLADYLKSKRRGERPHDILLASL
RGRKNDVEK

UMA1
MASASNAAGTSASAREYHANPTWKSKKEAQQVDYPSPEAFQSVMKRIERLPGL
VSPNEIDRLRSQLASVAEGNAFLLQTGDCAELFDYCNPEQIQAKLKLSLLMSLIL
LWGARKPVVRIGRIAGQYAKPRSKPTEIITKIDAATGVASQHEIMSFRGDNINAF
SAEPASGRIPDPERLLQAYFHSAATLNHIRSELASGLADLHAPRQWSFQHVQSR
ALQAEFESVVDSLTDALDFMKTIGADPSAFSSSSVLNSVDYFISHEGLSLAYEEA
LTRLIRKSTSVSPNPTRNAASSDSADDDLNGDVLERYDLSAHTIWLGDRTRQLD
GAHVDFFASIRNPVGVKVGPSMKPEELIQILDILNPDIPIPKATSSARASEWSTPS
AAAPEVVYGKEKGRVMVIIRLGAAKVASHLPPLLEAIAASNHKDSIILLCDPMH
GNTQTSPYPPASSDPNAQPLKTRSFGDIISEILSFLEIISADFPTLHLGGVHLELTG
DKDVTECFGGSMRLDPKHLERGYKSHCDPRLNFEQSLDVAFLLSHYFRNQRLG
NKRHAPAKTEDEIKKQQAIIASGSGDELLAELICGITNRAQMS

MGL1
MPSDSWRSKSIAQDVPYDDPKQLENVTKQLSSLPGLVTPAEIDRLHEQLARVAD
GKAFLLQCGDCAELFSYCNANQIEAKIKLILLMSLIIIYGARLPVVRVGRMAGQY
AKPRSKPYETITLPSGETKEVLSFRGDVVNSAEVDHRSPDPERLLSAYFHAAATL
NYTRGCLASGIADLHTPGTWRFHHVQSKSLQQEFERIADAISDALDFMKTVGA
DPTGDSNVLNTVDFFTSHESLILEFEHALTRDTPQGSYDLSAHTVWLGDRTRQP
DGAHVEFARHVRNPIGVKVGPSMKPEELITLLDTVNPHVQKGHVTLITRYGESK
VKDLLPEHIKAVQNSKHAKAVIWCCDPMHGNTVTSPSDPKLKTRMFSAVVSEL
TSCLQIHASLGSVMGGVHLELTGDEGVTECMGGSMELSDEDLKRCYLTHCDPR
LNYEQSLDIAFLISDVLKSQRRGIPVNNALSQALLRSNRVEGNP

SMT1
MSLSAVPPVPEPAATAAGAAWSPESWRGKTALQMPTYPDPVALDAALHELKR
LPPLVTSWEILALKQQLAEAQEGKRFLLQGGDCAENFSDCESGTISNRLKVLLQ
MSLVLVHGLRQPVIRVGRFAGQYAKPRSADTETRDGVTLPSYRGDVINAPAFTE
AARLPDPKRMLQAHAHSAMTMNFVRALIDGGFADLHHPEYWNLEWVSHSPLA
ADYQKMVASIGDAVHFMETLAGARVHNLNRIDFYTSHEALLLPYEQALTRQVP
RQQGWLNLSTHYPWIGMRTAALDGAHVEYLRGVRNPIAIKVGPSVTPDQLLRL
IDVLNPHDEPGRLSFIHRMGAAQIAEKLPPLLDAVKRDGRRVLWVCDAMHGNT
ESTANGFKTRRFDNVRGEVEMSFDLHAAAGTRLGGVHLELTGEDVTECTGGAR
ELTERDLERAYRSTVDPRLNYEQSLEIAMAIVRKQEQVR

BGL1
MECNLSSARVAENWTPASWKTKPALQQPVYRDAAALDRALTQLQALPPLVAA
WEVLTLKRKLVDAAEGRCFLLQGGDCAESFADCTAPIIANRLKVLMQMSLVLV
HGLMLPVVRVGRFAGQYAKPRSADTETRDGVTLPCYRGDIVNARAFTKDAREP
DPQRLLEAHSRSALTMNFVRALSDAGFADLHHPEHWDLAWASHSPLYGEYRK
MTHSIGESLRFMEALAGEPSDSDARVDFFTSHEVLLLNYEEAMARQAHRHWG
WFNLSTHFPWIGMRTAQLDGAHVEYCSGIRNPIGLKIGPGLSTDQLLRTIDALNP
TNEAGRLTLITRMGAAKIEAELPKHLHAVKTEGRRVLWCCDPMHGNGETTPGG
VKTRRFENINAELEAAFDIHAACGTHLGGVHLELTGEDVTECLGGARNLTEADL
ARAYKSTVDPRLNYEQSLEIAMLIVRKLGTTPVAEIGAERHTAYA

SRU1
MSEPTIQDWSLHSWREKDALQQPTYPDEAALEEKLDHVAALPPLVTSWEVENL
KAEIARAAQGDRFLLQGGECAESFGNCRADVITGRLKILMQMSLVLTYGLNTSI
VRVGRFAGQYAKPRSSDTETHDGTTLPSYRGDIINGPEFSAEARRPDPERMVEA
YSSSSLTLNLVRALAEGGFADLRHPEYWDLDFMDHSPLAEEYHALVDAIGDTID
FVEAVTEQELDSLESVTFYTSHEALLLPYEEALSRSVPHKDGIYNLGTHLPWVG
KRTNQVDNAHVEYARGIENPVGLKVGPDMTPSRLKTLVRTLDPEDEPGKLTLIS
RLGADTIGDRLPALVEAVQATGQSVLWIADPMHGNTETTDDGTKTRHFDNILG
ELEQALDVHAAEGSHLGGVHFELTGRDVTECIGGARGLSEADLGRAYESRVDP
RLNYEQSLEMAFSIVRKYRQLHG

SCL1
MYVPAQPWSPDSWRERFNAAPLPYDDPDAVQVAVEKLRALPPLVTSWEVERL
KGYLAEAQLGRRFVLQGGDCAETLDDCRPNIVTNKLKILLQMSLVLVHGARRP
VVRLGRFAGQYAKPRSSPVETRDGVSLPSYFGDLINRAPFDATARRPDPNLMIA
GYQHAAITLNFIRSLSEGGFADLHHPEYWDLSFLERADLPGELRERYQRMTEQV
GEALRFMEALAEVPVGELSRVEFFTSHEGLNLHYESAQTRKVARRPGFYDLTTH
LPWIGERTRQPDGAHVEFFSGVENPLGIKLGPKTLRDDVIRLLDRLNPADEPGKI
VLIVRMGAGRVRDGLPPLIEAVRSAGRKVLWMSDPMHGNTLTTSSGIKTRNFE
DILTEIEANFEVHESMGSNLGGVHFELTGDDVTECVGAGLTESDLDQRYASLCD
PRLNYRQSLEMAFRIAHRMAESSRRARG

SUS1

MGHRAPPHQRWLACDVISCGLPMWVPDSWRQRTALQQPEYADAAEVERVLA
EIRTLPPLVTSWEILHLREQLAEAAEGRQFVLQAGDCAERFVDCTPVRITNTLKV
LLQMSLVLVIGARRPVIRIGRFAGQYAKPRSTNDEVRDGITLPSYRGDNINRPEF
TAEARRPDPQLLLRGYERASLTLNFVRALVKGGFADLHHPEYFDLDWAKDSPL
AHEYHRMVQTIKDSLQFVENVLGVRAGDTDKIDFYTAHEALHLGYEAAQTRR
VPRRPGYFNLMTHFPWVGLRTNDPDGAHLEYFRGIENPIGIKVGGSATREQVAR
WFEMLDPERSPGRLTLIHRFGAGKISDALPRLIEHVRAEGGKPVWVCDPMHGN
THMTAGGIKTRNFEDIYAEVEKAFDIHASLGQKTGQKLGGVHIELTGENVTECV
GGARGPNEEGLARAYESEVDPRLNYEQSMELAFLIARKMKNGNG

MTU1
MNWTVDIPIDQLPSLPPLPTDLRTRLDAALAKPAAQQPTWPADQALAMRTVLE
SVPPVTVPSEIVRLQEQLAQVAKGEAFLLQGGDCAETFMDNTEPHIRGNVRALL
QMAVVLTYGASMPVVKVARIAGQYAKPRSADIDALGLRSYRGDMINGFAPDA
AAREHDPSRLVRAYANASAAMNLVRALTSSGLASLHLVHDWNREFVRTSPAG
ARYEALATEIDRGLRFMSACGVADRNLQTAEIYASHEALVLDYERAMLRLSDG
DDGEPQLFDLSAHTVWIGERTRQIDGAHIAFAQVIANPVGVKLGPNMTPELAVE
YVERLDPHNKPGRLTLVSRMGNHKVRDLLPPIVEKVQATGHQVIWQCDPMHG
NTHESSTGFKTRHFDRIVDEVQGFFEVHRALGTHPGGIHVEITGENVTECLGGA
QDISETDLAGRYETACDPRLNTQQSLELAFLVAEMLRD

SAQ1
MRHEWHQLSHPAVGSPVLQTSRPTVDSVEDAALGLDRWRDLPRAQTPPWPDD
AQVAEVCKVLENVPSVVAPYEVDQLRQRLALVCEGKAFLLQGGDCAETFADN
TESHLLANARTLLQMAIVLTYGASLPVVKVARVAGQYTKPRSLPTDARGLPAY
RGDMINSLDATPEARVADPQRMIRAYANSAAAMNMLRAYLAGGLADLHAVH
DWNKGFVRNSPAGERYEAIAREIDRALAFIRACGMTEDEALRTVTLYCSHEALA
LEYDRALTRVSDNRAYGLSGHLLWVGERTRQLDGAHVDFISRIANPIGVKLGPT
TKPDDVIELCEKLNPDNVPGRLTLISRMGNHRVRDALPPIVAKVTAAGAKVVW
QCDPMHGNTHESSNGYKTRHFDRIVDEVLGYFEVHRGLETHPGGLHVELTGED
VTECLGGAQGIDDIDLPDRYETACDPRLNTQQSLELAFLVAEMLRG

STP1
MYGTRMRPGVERSIRHSPPATNDGDRELCLDRWRELPRRQVPPWPDPAEVAAV
CATLGKMPPIVTPYEVDELRHRLAEVCEGRAFLLQGGDCAETFTGNTESHLLGT
TRTLLQMAMAITYGGSVPVVKVARLAGQYGKPRSSATDSLGLPAYRGDIINAR
HPAESARAADPQRMIDAYANSAVAMNLIRAYPPDDLTDLEELYDDTYDLIRASP
AGARYQVISGEIDRARGFVRAWGPSERHALRESKVYCSHEALVLEYDRALTRIN
DGRAYALSGHFLWVGERTRQLDHAHVDFVARIANPIGVKLGPAASPHAAIELC
ERLNPENLPGRLTLISRMGNRQVRDVFPAIVDKVTAAGAKVVWQCDPMHGNT
EQSSHGFKTRRLDRVVDELLGYFDVHRSLGTHPGGVHVELTGENVTECLDGIR
GVEDQHLPDRYETACDPRLNMRQSLELALLVAEILRG

FRE1
MSNDLDIWRSLPARQQPSWPDGEELAAAFAELSALPPLVTAPEVRSLTDRLAM
VARGEAFLLQGGDCAETFAANTADKIRDKVKTLLQMAVALTYGASTPVVKVA
RIAGQYAKPRSADIEASTGLPSYRGDAVNDIAPNAQARRPNPRRMVDAYHQSA
VALNLVRAFATGGFADLSKVHEWNKAFVRDSAAGRRYELMAVDIERALAFMA
ACGIDLDRTAALTGVEMFTSHEGLLMEYERALTRTEESTGEVYDLSAHMIWIGE
RTRDLDGAHVDFLSRVGNPIGCKIGPTATPDEVVALTERLNPDHIPGRLTLIARM
GAKRVRDALPPIIDKVNAAGHPVVWSCDPMHGNTRDVGGVKTRHFDDVLDEV
FGFFEVHKGLGTHPGGLHIELTGENVTECLGGAEMIGEADLGGRYETACDPRLN
TGQALELAFLVAESLQQARAERDTHTR

ACE1
MLLSDPPLESWRQLPAAQQPIWPDPDELAAVRSQLTALPPLVVPEECDVLRQRL
AEVARGRAFLLQGGDCAETFADATVDRIRSKIKILLQMAVVLTYGASVPVVKV

GRLAGQYAKPRSRDVEVVDGVELPSYRGDAVNSIERTLAARVPDPRRLLEAYH
RAAMTLNLIRAHATGGFADLRKVHAWNSDFVRRSPAGERYESIATDIDRALRF
MEACGFDLENSEVAHQVELYASHEALLLDYESALTRYDTERRAYYDLSAHMV
WIGERTRDVGGAHIEFARYIANPIGVKLGPTTTPETAVALVEKLDPERRPGRLTF
ITRMGAGRIRDVLPPIVEKVTAAGAEVIWSCDPMHGNTIESSGGLKTRHFDDIVE
EVMGFFEVHQALGTIPGGLHVELTGEDVTECLGGADDIADADLADRYETACDP
RLNAGQSLELAFLAASMLQDIRRRADESANPREAS

SGR1
MLVTVNANTSVAGGNTWRDLPAAQQPEYPDSEALRDVLADLASYPPLVFAGE
CDQLRARLGAVAKGEAFLLQGGDCAEAFDAVSAEHIRAKLKTLLQMSAVLTY
AASVPVVKVGRIAGQYSKPRSKSTETRDGVTLPTYRGDSVNGFAFTEEARVPDP
QRLKQMYHASSSTLNLVRAFTTGGYADLRQVHAWNQDFVKSSPSGQRYEALA
REIDNALNFMKACGTDPAEFKAVEFYASHEALLLDYESALTRTDSRTGELYDTS
AHMVWIGERTRQMDGAHIEFAARIRNPIGIKLGPTTTVDEALGYIDRLDPEREPG
RLTFVVRMGADKVRDKLPELVEKVTASGATVAWVTDPMHGNTFEAASGHKT
RRFDDVLDEVKGFFEVHKALGTHPGGIHVELTGDDVTECVGGGHEIFVDDLHQ
RYETACDPRLNRSQSLDLAFLVAEMYRDQ

SEN2
MRNLDAWRGLPAAQQPEWHDPAELSSVLDELSQLPPLVFAGECDQLKDRLAA
AQRGESFVLQGGDCAETFAGATASAVQNKLKTLLQMAVVLTYAAQVPVVKIG
RMAGQFAKPRSSPDETSDGVTLPAYRGDAVNGFEFTAESRRPDPRRLIRAYHSS
ATILNLCRAFANGGYADLRQVHAWNQDFVATSPAGKDYEHLASEIDRALAFM
RACKAEPAEVHTVEFWSSHEALLLDYESALTRVDPRTSWPYAVSAHMVWIGER
TRQLDHAHLEWASRIRNPLGVKLGPSTTPDDALAIIERLNPSAEPGRLTFVTRMG
ADRVRDTLPLLVEKVTASGAPVTWVCDPMHGNTFRSPSGYKTRRVDDIFDEVT
GFFEVHRALRTHPGGIHIEFTGDEVTECVGGGHGLLEADLHHRYETVCDPRLNR
RQSLDLAFKVAELYRLST

SEN3
MPAAQQPDWPDPVGLRRVVDELRAAPPLVFPSECDRLRSRLAVVARGGGFLLQ
GGDCAETFDGVSPAQVRAKLRTLAQMALVIGYACAAPVVKVGRIAGQYSKPRS
SPTETRDGRTLPSYRGDSVNGPEFGEQARTPDPERLLRMYHSSMATLNVLRAFG
GGGRSDLRRVHGWNEEFVSASPSGKHYQRLWEEIGAALGFMRSLGAAPETVGE
TEFFTSHEALVLEYESALTHGGGRPGEAYDLSGHMVWIGERTRRIDGAHVEFAS
RIRNPIGIKLSAAAEPDEVLALIDKLDPHREPGRLTLITRMGAQRVRAELPALVE
KVTAAGAPVVWVCDPMHGNTFEAANGYKTRRFDDVLDELRGFLEVHRALGT
HPGGVHLELTGDDVTECLGGGDDIGPDDLPERYRTACDPRLNRSQALDLAFRL
AEMYRGA

PAC1
MAGSRIPAMSQSVPSLEALNALPRVQQPSYPDPQATVEVVNQLRSLPPLVFAGE
CDDLRAYLAAVANRKAFLLQGGDCAETFAGVNADNIKGKLRVLLSMAVVMT
YAGQLPVVKVGRIAGQYAKPRSKDTETRRGLTLPSYRGDAVNGFEFTPEAREH
DPKRLIGVYNASAATLNLVRAFATGGFADLRGIHAWNADFVRNSNVETRYEAL
AAEIERALVFMMACGVEGDQLSTVDFYASHEALLIDYEHAMTRIDSRSQLPYD
CSGHLLWIGERTRQIDGAHVEFLRHVRNPLGVKLGPTTTGEDAVAIADALDPDH
EPGRLTFITRMGSKAVRTNLPRVIEAVEATGRKVVWSCDPMHGNTFETNDGYK
TRSFADMCDEVNGFFDVHEELGTWPGGVHIELTGDDVTECLGGVDKLAESDLT
NRYETACDPRLNRNQSLELAFIIAERLADSRIKRDAVSPLARFRSIDL

NCA1
MSTIPDLATLHAMGAAQQPSYPDRAAVDSAVQRLRTAPPLVFAGECDDLKGKI
AQVARGEAFLLQGGDCAETFAGVTADNVRNKLRVLLQMAVVLTYAASVPVV
KVGRIAGQYAKPRSSDFETRDGVTLPAYRGDAVNGYDFTAESRVPDPQRLVDV
YNSSAATLNLVRAFVTGGYADLRQVHTWNTDFVRESPVGQQYEAMANEIERA

LTFMRAIGADPDEFHRVDFHSSHEALVLEYEQALTRIDSRTSTPYDVSGHFLWIG
ERTRQLDGAHVELLSHIRNPIGVKLGPTTTPDDALALAAKLNPDNEPGRLTFITR
FGAGKIRDGLPTLVEKVTAAGLEVAWVCDPMHGNTFEASSGYKTRRFGDVIDE
VQGFFDVHRSLGTWPGGLHVELTGDDVTECVGGGEDLMEVDLGNRYESVCDP
RLNRVQSLELAFLVAEMLRQA

TWS1
MPGSNRVNPFIKNSVDVSLETINAWRGLPIEQQPEWPDSVLLSEVSCTLSQLPPL
VFAGEVDKLRLHLQEAAEGRAFILQGGDCAESFRDITADNISRKLKTILQMAAIL
TYAAGVPIIKMARMAGQYAKPRSEQFELVDGRSLPVYRGDMVNGHELHDRVP
DPKRMLKAYELSAATLNLIRAFTRGGFADLRALHVWNRGFAESCENEQYEELA
RKIDKAIAFMNACGVDCSNLTQTEVFVCHEGLLLDYEHALTRCDSRTGFFYDTS
AHFLWIGERTRQIDGAHIDFFSKIENPIGVKLSSKVTARQVLELANRLDPGRTPG
RLTFITRMGAGNIRQCLPGIIEAVRDGGFNPAWVCDPMHGNTITSSSGYKTRRF
MDILDEVRGFFDVHRSVGTHPGGLHVELTGDDVTECTGGFDSISDDDLKDRYES
LCDPRLNHRQSLELAFFLADCMK

LXX1
MMEGLDYWRTLPIKQQPEWPDKDAVTAVSAELATLPPLVFAGEVDQLRTRLA
RAAEGNAFLLQGGDCAEKFADATADQIRNRVKTVLQMAVVLTYGASVPVIKM
GRMAGQFAKPRSSDTETRGEVTLPAYRGDIVNGYDFTPESRQADPRRLVQGYH
MAASTLNLIRAFTQGGFADLRQVHSWNRGFAANPANQRYEGLAREIDRAIKFM
EAAGADFDELKRVEFYSSHEALLMDYERPMTRIDSRTGTPYNTSAHFVWIGERT
RDLDGAHVDFLSRVRNPIGVKLGPTTSPDDMLRLIDKLDPEREPGRLTFVTRMG
AGRVREALPPLLEAIKRSDACPLWVTDPMHGNGLTTPTGYKTRRFEDVVDEVK
GFFEAHRAAGTNPGGIHVELTGDDVTECLGGSEHIDEATLATRYESLCDPRLNH
MQSLELAFLVAEELSQP

RSA1
MQDQTNSPFMTLDDSAVRLGQAGAANYPGLDAWRSLPISQQPTWQDEAVYQS
SIKELSSLPPLVFAGEVDILRDRLAAAAQGKAFLLQGGDCAETFDGATADKISAR
VKTILQMAVVLTYGASLPVIKMGRMAGQFAKPRSSNDETRDGVTLPAYRGDM
VNDYEFTPESRGHDAQRLVRAYHTSASTLNLIRAFTQGGFADLRLVHHWNKGF
TQNPAHARYESLAGEIDRAIRFMEACGSDFEALKRTEFFASHEALLLDYERSLTR
IDSRTGTPYATSGHFLWIGERTRQLDAAHVDYLSQVRNPIGVKLGPSTSGDDAL
RLIDKLDPNREPGRITFITRMGAKNIREKLPALVQKVTDSGAQVLWVTDPMHGN
TVTSPNGYKTRNFDDVMDEVRGFFEVHQALGTFPGGLHVEMTGDDVAECLGG
ADPIDQEAFLTRYESVCDPRLNHMQSLEMAFLVSGALAQR

KRH1
MSAETLSDQFRSVMTPSPSAADYPGLDEWRDKTIGQHPQWAQHPDFAGSIQEL
NAVPPLVFAGEVDRLRSRLAQVANGEAFFLQGGDCAETFAGATADKISGRVRT
LLQMAVVLTYGASLPVVKMGRMAGQFSKPRSSNEETRDGVTLPSFRGEMVNG
FDFTEQSRVHDPKRMVRGYHTSASTLNLIRAFTQGGFADLRRVHAWNMGFAA
NPAYARYESMAREIDHAVKFMAACGSDFDALKTTEFFAGHEALLLDYERALTR
IDSRTNLPYGTSAHFLWIGERTRDLDGAHVDLLSRLRNPIGVKLGPGTTGEDAL
RLIDALDPQREPGRLTFITRMGAQNIREKLPPVIQAVEASGAKPVWVTDPMHGN
TVSLPSGYKTRKFEDVVDEVRGFFEVHRELGTFPGGIHVEMTGDDVAECLGGS
DPVAESSFAERYETLCDPRLNHMQSLEIAFLVAEYLSAQN

MLU1
MTEDPREKILAFGTAISEGAADIPGLDHWRSLPIAQAPSWADEADHHAAVAELS
ALPPLVFAGEVDLLRARLADVAQGRAFLLQGGDCAETFAGSTANRISARVKTIL
QMAAVLTYGASMPVVKMGRIAGQFAKPRSSDTETRDGVTLPSFRGEIVNGYEF
TEQARRHDARRMVQAYHTSASTLNLVRAFTQGGFADLRSVHQWNQGFMANP
AYAQYEEMAAEIDRAVRFMNAAGVDFEAMRRTEFYSAHEALLLDYERALTRV
DSRSGDPCATSAHFLWIGERTRQVDGAHVDFLSRVRNPIGVKLGPTTTVADALE

LAERLDPEREPGRLTFITRMGAERIRDVLPGIVEGVRDAGLEPVWVTDPMHGNT
VTAANGYKTRRFEDVMDEVSGFFEVHRELGTHPGGMHVELTGDDVAECLGGS
DPINEADFDARYETLCDPRLNHQQSLEMAFKVTEALSHPERRAPRA

BCV1
MTIPADPQVLAGLDAWRRLPAVQQPEWPDAAVLASIGAELAAAPPLVFAGEAD
ALREKLAAASRGEAFLLQGGDCAETFAESSADNIRNKVKTILQMAVVLTYGAS
LPIVKMGRMAGQYAKPRSTGTETRDGVTLPAYRGDTINGHEFTAASRTPDPAR
LLKAYHTSAATLNLIRAFTMGGFADLRRVHEWNRGFMTNPSYKQYDELASEID
RAIRFMAAAGADFDALRTVEFFSSHEGLLLDYERPLTRIDSRTGLPYDCSAHFL
WIGERTRQLDGAHVDFFSRVHNPIGVKLGPTTSGDDALALMDALNPDGVAGRL
TFITRMGAGRIRDALPGVVEKVAADGRPVVWVCDPMHGNTITSANGYKTRRFT
DVIDEVRGFFEVHAATGTVPGGLHVELTGDDVTEVLGGAEEIDDVGLARRYTT
LVDPRLNHQQSLEMAFLVAEMLRERPTA

KRA1
MSTPTTTSPLSSGQSSITWPDLPAAQQPQWPDAAALAEVQAELASYPPLVFAGE
ADQLRAQLAKAARGEAFLLQGGDCAETFAGATADGIRRRLKTMLQMSAVLTY
GASLPVIKVGRMAGQFSKPRSSNDETRDGVTLPAYRGDAVNDYAFTPESRVPD
PQRLLRAYHTSSATLNLIRAFTTGGYADLRKVHEWNRGFAATAANAKYDSLAR
DIDKAMRFMAACGADFDALQTVQFWSSHEALLLDYERPLTRTLDEEPPLAGAA
PDDVRQGGPYDVAAHMVWIGERTRQLDGAHVDFAARIRNPIGIKLGPKTSVDD
ALALIDKVDPYREPGRLTFITRMGASTIREALPALVEGVTRSGAQVTWVCDPMH
GNTITSTSGYKTRRFDDVVDEVRGFFEVHRALGTVPGGLHVELTGDDVTECLG
GAYDIDDAALATRYESLCDPRLNHQQSLELAFLAAEMLAER

SCO2
MTASDTYTQPRTAPAGTGTAAPGPTGPAADAGRGGWTPGSWRSRPAAQQPEW
PDPGELRGVEDTLALRPPLVLPDEILDLRRSLAQVAAGEGFLLQAGDCAERFGS
CTEAGVRGKLRVILQVAILLTYGSGLPVVKVGRIAGQFGKPRSRPTEAVDGVEL
PVYRGDIVNGPEFTAEARRPDAGRLLSAYHHASAALNVLRALTLGGYADLGQV
HEWNQEFVRRSPAGQRYEKAADDITWALRFMSACGLDTRSQAALHQVQLYTS
HEALLPQYEQALIRYDEKRRGWFDTSAHLLWIGDRTRRVDGAHVELLAGVDNP
VGVKVGPTMGVDDLRELCERLDPDRAPGRLVLISRLGAGRGAELLPPLLRAVR
DAGHAPVWACDPMHGNTFVSASGYKTRRLSDITTEVAEFFAVHREEGLHPGGI
HLELTGDDVTECLGGDLDEVLDTHLASRYETACDPRLNAAQSIELAFGVARLLR
ELRGGA

SEN4
MPRTQQAALSAVPAAQQPVWPDPAALRTALAELAVQPHLVSPRECDLLRDQL
AFVAGGGAFLLQGGDCAETFAGAASEAVGRKVRLLGQMAEILAERMGLPVVT
VGRIAGQYAKPRSWPTESHLGQTLPSYRGDAVNGLAFTPAARTPDPRRLLQAY
RTAAGTLAVLRKQDEPASPGDSLLPHPRTSDGGPGGNDFFTSHEALLLPYEEAL
CRTDPLTGQRYGTSGHLLWIGERTRNLDGAHVRFAAGISNPVGVKLGPAVNPD
DVLALVDKLDPQRQPGRLSLIIRMGARKIRDLLPPLVEKVAAEHAPVTWVCDP
MHGNTFTAASGHKTRLFDDVFDEVAGFFEVHRTLGTHPGGLHLELTGEAVTEC
VGGTPAVLMEDLDRRYETACDPRLNADQARELAVAVGGLAARP

PTM1
MNQDKEWSLDSWKNYSAVQLPIYSDVQLYQDTLQKIKKLPEILNFEDINAFKTQ
MIKVSKGEKFVLQLGDCAEVFDECTEEHWKEKFSFYDRMGQILDAIVIGRTCGQ
FAKPRSQLLEKDGTLNYRGDLINSLSRDEREPDPSRLLLGAHYTKKGLESLKQY
EGQQIWVSHECLHLGYESAFVKQNEDQQYYLSNTHFPWIGDRTRLHDHAHSNF
IRGIYNPVGIKIGQTINKTEFINVFKLINPKNEDGRSFAIIRLGKNKLDKLKDIINW
KQEEKLNISFFLDPMHGNNQDKSGCKIRKIEDIMYEIQQFFTIMEQQNEVPAGLH
LECTPYDVTECIENDEDINPKKYTTACDPRLNFRQTRKIIIFVNTLLKKLRNKQ

PTM3
MNQDKEWSIDGWKKHSALQLPIYADAQLYQDTLQKIKNLPGILNFEDINAFKN
QMIRVSKGEKFVLQLGDCAEVFDECTEKHWKEKFSFYDRMGQILDAIVIGRTC
GQFAKPRSQLLEKDGTLNYRGDLINSLSRDERDPDPSRLLLGAHYTKKGIETLK
QYEGSQIWVSHECLHLGYESAFVKQNQDKQYYLSNTHFPWIGDRTRLHDHAHS
NFIKGIYNPVGIKIGQTINKTEFVNVFKLINPNNEDGRAFAIIRLGKNKLDKLKDII
QWKQEEKLNISFFLDPMHGNNLDKGGRKIRKIEDIMYEIQQFFAILEQENESPAG
LHLECTPYDVTECIENDEDINPIKYTTACDPRLNSRQTRKIIIFVNTLLKKLRNKQ

PTM2
MNQEKEWNLDSWKKYPALQQPLYQDLHFYQETLQKLRKLPGILNFEEINKFKN
EMIKVSQGDKFVLQLGDCAEVYNECTEEHWKEKFSFYDSMGQILNAIVIGRTC
GQFAKPRTQLHEKDGTLNYRGDLINSLSRDERDPDPLRLLLGAHYSKKGIDTLR
KYEGKQLWVSHECLHIGYESAFARQHEEDHYYLSNTHFPWIGDRTRLCEHAHS
NFAKGIYNPIGIKIGSSINKNEFVNLIKLLNPKNEDGRVFAIIRLGKNKLDKLKDII
NWKHEEKLNVSFFLDPMHGNSLEQNGYKIRRIEDIMFEIQQFFNILEQSNEKPAG
LHLECTPYDVSECVENDEEIDPKKYTTACDARLNFRQTRKIVIFVNTLLKKLKK
KQQ

TET1
MIESSVLLESELQNKWDGSEIRFSIGAKTNPVNSKQHVLQQKRSFIEDEGTGSET
SEEISHNEIPETWLPSSWRSYPALQQPVYENTNLLSQQLQMINEMPSIVSVNEIES
LQADLRQVETGSAFFIQMGDCAERFAEMNNSFMEKKIKAITAASRIFANMLNK
KCVKIGRIAGQFAKPRSQEFEKLPDGTLQYNYRGDNVNSLESRKPDPLRLSQGY
FHSVAAMNFIRAQQKHIMHNKLNIQQLLEEKSYPQKQQISAEQQQKLNDNQIEA
IENVNLNNDDLYISHEGLLLNYEEAMTKKYNNKYYNTSTHLLWMGERTRQLD
GAHVNFFSGIQNPIAVKVGPTTDGKILKQLVNKLNPQNLQGRLIIICRMGNKNIS
NVLPTLVQAKKEYNLNFIWSCDPMHGNTFSTNENIKTRSVKDILEELTKFATILH
ENKENIGGIHLETSPNQVTECLGLEIEDQDLNICYTSACDPRLNISQTLYITYQFA
LLLQSLYNKKNTGYTPELDN

References

1. Wang, K. K., Villalobo, A., and Roufogalis, B. D. (1989) Calmodulin-binding proteins as calpain substrates, *Biochem. J. 262*, 693-706.
2. Croall, D. E., and Demartino, G. N. (1991) Calcium-activated neutral protease (calpain) system - structure, function, and regulation, *Physiol. Rev. 71*, 813-847.
3. Greenwood, A. F., and Jope, R. S. (1994) Brain G-protein proteolysis by calpain - enhancement by lithium, *Brain Res. 636*, 320-326.
4. Bi, X. N., Tocco, G., and Baudry, M. (1994) Calpain-mediated regulation of ampa receptors in adult rat brain, *Neuroreport 6*, 61-64.
5. Hirai, S., Kawasaki, H., Yaniv, M., and Suzuki, K. (1991) Degradation of transcription factors, C-jun and C-fos, by calpain, *FEBS Lett. 287*, 57-61.
6. Watt, F., and Molloy, P. L. (1993) Specific cleavage of transcription factors by the thiol protease, m-calpain, *Nucl. Acids Res. 21*, 5092-5100.
7. Goll, D. E., Thompson, V. F., Li, H., Wei, W. E. I., and Cong, J. (2003) The calpain system, *Physiol. Rev. 83*, 731-801.
8. Donkor, I. O. (2000) A survey of calpain inhibitors, *Curr. Med. Chem. 7*, 1171-1188.
9. Zatz, M., and Starling, A. (2005) Calpains and disease, *New Engl. J. Med. 352*, 2413-2423.
10. Strobl, S., Fernandez-Catalan, C., Braun, M., Huber, R., Masumoto, H., Nakagawa, K., Irie, A., Sorimachi, H., Bourenkow, G., Bartunik, H., Suzuki, K., and Bode, W. (2000) The crystal structure of calcium-free human m-calpain suggests an electrostatic switch mechanism for activation by calcium, *Proc. Natl. Acad. Sci. U. S. A. 97*, 588-592.
11. Alexa, A., Bozoky, Z., Farkas, A., Tompa, P., and Friedrich, P. (2004) Contribution of distinct structural elements to activation of calpain by Ca^{2+} Ions, *J. Biol. Chem. 279*, 20118-20126.
12. Bozoky, Z., Alexa, A., Tompa, P., and Friedrich, P. (2005) Multiple interactions of the 'transducer' govern its function in calpain activation by Ca^{2+}, *Biochem. J. 388*, 741-744.
13. Takano, E., and Maki, M. (1999) Structure of calpastatin and its inhibitory control of calpain, In *CALPAIN: Pharmacology and Toxicology of Calcium-Dependent Protease* (Wang, K. K. W., and Yuen, P.-w., Eds.), pp 25-50, Taylor &Francis.
14. Asada, K., Ishino, Y., Shimada, M., Shimojo, T., Endo, M., Kimizuka, F., Kato, I., Maki, M., Hatanaka, M., and Murachi, T. (1989) cDNA cloning of human calpastatin: sequence homology among human, pig, and rabbit calpastatins, *J. Enzym. Inhib. 3*, 49-56.
15. Lee, W. J., Ma, H., Takano, E., Yang, H. Q., Hatanaka, M., and Maki, M. (1992) Molecular diversity in amino-terminal domains of human calpastatin by exon skipping, *J. Biol. Chem. 267*, 8437-8442.

16. Maki, M., Takano, E., Osawa, T., Ooi, T., Murachi, T., and Hatanaka, M. (1988) Analysis of structure-function relationship of pig calpastatin by expression of mutated cDNAs in *Escherichia coli*, *J. Biol. Chem. 263*, 10254-10261.
17. Emori, Y., Kawasaki, H., Imajoh, S., Imahori, K., and Suzuki, K. (1987) Endogenous inhibitor for calcium-dependent cysteine protease contains 4 internal repeats that could be responsible for its multiple reactive sites, *Proc. Natl. Acad. Sci. U. S. A. 84*, 3590-3594.
18. Killefer, J., and Koohmaraie, M. (1994) Bovine skeletal-muscle calpastatin - cloning, sequence-analysis, and steady-state messenger-RNA expression, *J. Anim. Sci. 72*, 606-614.
19. Melloni, E., Tullio, R. D., Averna, M., Tedesco, I., Salamino, F., Sparatore, B., and Pontremoli, S. (1998) Properties of calpastatin forms in rat brain, *FEBS Lett. 431*, 55-58.
20. Averna, M., De Tullio, R., Salamino, F., Melloni, E., and Pontremoli, S. (1999) Phosphorylation of rat brain calpastatins by protein kinase C, *FEBS Lett. 450*, 13-16.
21. Melloni, E., Averna, M., Stifanese, R., De Tullio, R., Defranchi, E., Salamino, F., and Pontremoli, S. (2006) Association of calpastatin with inactive calpain - a novel mechanism to control the activation of the protease?, *J. Biol. Chem. 281*, 24945-24954.
22. Kiss, R., Kovacs, D., Tompa, P., and Perczel, A. (2008) Local structural preferences of calpastatin, the intrinsically unstructured protein inhibitor of calpain, *Biochemistry 47*, 6936-6945.
23. Maki, M., Takano, E., Mori, H., Sato, A., Murachi, T., and Hatanaka, M. (1987) All four internally repetitive domains of pig calpastatin possess inhibitory activities against calpains I and II, *FEBS Lett. 223*, 174-180.
24. Emori, Y., Kawasaki, H., Imajoh, S., Minami, Y., and Suzuki, K. (1988) All four repeating domains of the endogenous inhibitor for calcium-dependent protease independently retain inhibitory activity. Expression of the cDNA fragments in *Escherichia coli.* , *J. Biol. Chem. 263*, 2364-2370.
25. Hanna, R. A., Garcia-Diaz, B. E., and Davies, P. L. (2007) Calpastatin simultaneously binds four calpains with different kinetic constants, *FEBS Lett. 581*, 2894-2898.
26. Todd, B., Moore, D., Deivanayagam, C. C. S., Lin, G. D., Chattopadhyay, D., Maki, M., Wang, K. K. W., and Narayana, S. V. L. (2003) A structural model for the inhibition of calpain by calpastatin: crystal structures of the native domain VI of calpain and its complexes with calpastatin peptide and a small molecule inhibitor, *J. Mol. Biol. 328*, 131-146.
27. Nishimura, T., and Goll, D. E. (1991) Binding of calpain fragments to calpastatin, *J. Biol. Chem. 266*, 11842-11850.
28. Ma, H., Yang, H. Q., Takano, E., Hatanaka, M., and Maki, M. (1994) Amino-terminal conserved region in proteinase inhibitor domain of calpastatin potentiates its calpain inhibitory activity by interacting with calmodulin-like domain of the proteinase, *J. Biol. Chem. 269*, 24430-24436.
29. Hanna, R. A., Campbell, R. L., and Davies, P. L. (2008)

Calcium-bound structure of calpain and its mechanism of inhibition by calpastatin, *Nat. Lett. 456*, 409-412.
30. Moldoveanu, T., Gehring, K., and Green, D. R. (2008) Concerted multi-pronged attack by calpastatin to occlude the catalytic cleft of heterodimeric calpains, *Nat. Lett. 456*, 404-408.
31. Ishima, R., Tamura, A., Akasaka, K., Hamaguchi, K., Makino, K., Murachi, T., Hatanaka, M., and Maki, M. (1991) Structure of the active 27-residue fragment of human calpastatin, *FEBS Lett. 294*, 64-66.
32. Betts, R., Weinsheimer, S., Blouse, G. E., and Anagli, J. (2003) Structural determinants of the calpain inhibitory activity of calpastatin peptide B27-WT, *J. Biol. Chem. 278*, 7800-7809.
33. Betts, R., and Anagli, J. (2004) The β- and γ-CH_2 of B27-WT's Leu11 and Ile18 side chains play a direct role in calpain inhibition, *Biochemistry 43*, 2596-2604.
34. Pfizer, J., Assfalg-Machleidt, I., Machleidt, W., and Schaschke, N. (2008) Inhibition of human μ-calpain by conformationally constrained calpastatin peptides, *Biol. Chem. 389*, 83-90.
35. Darden, T., York, D., and Pedersen, L. (1993) Particle mesh Ewald: an Nlog(N) method for Ewald sums in large systems, *J. Chem. Phys. 98*, 10089-10092.
36. Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L., and Schulten, K. (2005) Scalable molecular dynamics with NAMD, *J. Comput. Chem. 26*, 1781-1802.
37. Ewald, P. (1921) Die berechnung optischer und elektrostatischer Gitterpotentiale, *Ann. Phys 369*, 253–287.
38. MacroModel, version 9.1,Schrodinger, LLC, New York, NY, 2005.
39. Maestro, version 7.5, Schrodinger, LLC, New York, NY, 2005.
40. Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical integration of the Cartesian equation of motion of a system with constraints: molecular dynamics of N-alkanes., *J. Comput. Phys. 23*, 327.
41. MacKerell Jr., A. D., Bashford, D., Bellott, M., Dunbrack Jr., R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher III., W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wio´rkiewicz-Kuczera, J., Yin, D., and Karplus, M. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins, *J. Phys. Chem. B 102*, 3586-3616.
42. Shenkin, P. S., and McDonald, D. Q. (1994) Cluster analysis of molecular conformations, *J. Comput. Chem. 15*, 899-916.
43. Herrmann, K. M., and Weaver, L. M. (1999) The shikimate pathway, *Annu. Rev. Plant Physiol. Plant Molec. Biol. 50*, 473-503.
44. Herrmann, K. M. (1995) The shikimate pathway - early steps in the biosynthesis of aromatic compounds, *Plant Cell 7*, 907-919.
45. Bentley, R. (1990) The shikimate pathway - a metabolic tree with many branches, *Crit. Rev. Biochem. Mol. Biol. 25*, 307-384.
46. Walsh, C. T., Liu, J., Rusnak, F., and Sakaitani, M. (1990) Molecular

studies on the enzymes in chorismate metabolism and the enteribactin biosynthetic pathway, *Chem. Rev. 90*, 1105-1129.

47. Roberts, F., Roberts, C. W., Johnson, J. J., Kyle, D. E., Krell, T., Coggins, J. R., Coombs, G. H., Milhous, W. K., Tzipori, S., Ferguson, D. J. P., Chakrabarti, D., and McLeod, R. (1998) Evidence for the shikimate pathway in apicomplexan parasites, *Nature 393*, 801-805.
48. Campbell, S. A., Richards, T. A., Mui, E. J., Samuel, B. U., Coggins, J. R., McLeod, R., and Roberts, C. W. (2004) A complete shikimate pathway in *Toxoplasma gondii:* an ancient eukaryotic innovation, *Int. J. Parasitol. 34*, 5-13.
49. Parish, T., and Stoker, N. G. (2002) The common aromatic amino acid biosynthesis pathway is essential in *Mycobacterium tuberculosis*, *Microbiology 148*, 3069-3077.
50. Rattan, A., Kalia, A., and Ahmad, N. (1998) Multidrug-resistant *Mycobacterium tuberculosis*: molecular perspectives, *Emerg. Infect. Dis. 4*, 195-209.
51. Parrish, N. M., Dick, J. D., and Bishai, W. R. (1998) Mechanisms of latency in *Mycobacterium tuberculosis*, *Trends in microbiology 6*, 107-112.
52. Stokstad, E. (2000) Infectious disease: drug-resistant TB on the rise, *Science 287*, 2391.
53. Ducati, R. G., Basso, L. A., and Santos, D. S. (2007) Mycobacterial shikimate pathway enzymes as targets for drug design, *Curr. Drug Targets 8*, 423-435.
54. Walker, G. E., Dunbar, B., Hunter, I. S., Nimmo, H. G., and Coggins, J. R. (1996) Evidence for a novel class of microbial 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase in *Streptomyces coelicolor* A3(2), *Streptomyces rimosus* and *Neurospora crassa*, *Microbiology 142*, 1973-1982.
55. Jensen, R. A., Xie, G., Calhoun, D. H., and Bonner, C. A. (2002) The correct phylogenetic relationship of KdsA (3-deoxy-D-*manno*-octulosonate 8-phosphate synthase) with one of two independently evolved classes of AroA (3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase), *J. Mol. Evol. 54*, 416-423.
56. Subramaniam, P. S., Xie, G., Xia, T., and Jensen, R. A. (1998) Substrate ambiguity of 3-deoxy-D-*manno*-octulosonate 8-phosphate synthase from *Neisseria gonorrhoeae* in the context of its membership in a protein family containing a subset of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthases, *J. Bacteriol. 180*, 119-127.
57. Shumilin, I. A., Kretsinger, R. H., and Bauerle, R. H. (1999) Crystal structure of phenylalanine-regulated 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase from *Escherichia coli*, *Structure 7*, 865-875.
58. Wagner, T., Shumilin, I. A., Bauerle, R., and Kretsinger, R. H. (2000) Structure of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase from *Escherichia coli*: Comparison of the Mn^{2+}*2-phosphoglycolate and the Pb^{2+}*2-phosphoenolpyruvate complexes and implications for catalysis, *J. Mol. Biol. 301*, 389-399.

59. Shumilin, I. A., Bauerle, R., and Kretsinger, R. H. (2003) The high-resolution structure of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase reveals a twist in the plane of bound phosphoenolpyruvate, *Biochemistry 42*, 3766-3776.
60. Hartmann, M., Schneider, T. R., Pfeil, A., Heinrich, G., Lipscomb, W. N., and Braus, G. H. (2003) Evolution of feedback-inhibited β/α barrel isoenzymes by gene duplication and a single mutation, *Proc. Natl. Acad. Sci. U. S. A. 100*, 862-867.
61. König, V., Pfeil, A., Braus, G. H., and Schneider, T. R. (2004) Substrate and metal complexes of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase from *Saccharomyces cerevisiae* provide new insights into the catalytic mechanism, *J. Mol. Biol. 337*, 675-690.
62. Shumilin, I. A., Bauerle, R., Wu, J., Woodard, R. W., and Kretsinger, R. H. (2004) Crystal structure of the reaction complex of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase from *Thermotoga maritima* refines the catalytic mechanism and indicates a new mechanism of allosteric regulation, *J. Mol. Biol. 341*, 455-466.
63. Schofield, L. R., Anderson, B. F., Patchett, M. L., Norris, G. E., Jameson, G. B., and Parker, E. J. (2005) Substrate ambiguity and structure of *Pyrococcus furiosus* 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase: an ancestral 3-deoxyald-2-ulosonate-phosphate synthase?, *Biochemistry 44*, 11950-11962.
64. Shumilin, I. A., Zhao, C., Bauerle, R., and Kretsinger, R. H. (2002) Allosteric inhibition of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase alters the coordination of both substrates, *J. Mol. Biol. 320*, 1147-1156.
65. Helmstaedt, K., Strittmatter, A., Lipscomb, W. N., and Braus Gerhard, H. (2005) Evolution of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase-encoding genes in the yeast *Saccharomyces cerevisiae*, *Proc. Natl. Acad. Sci. U. S. A. 102*, 9784-9789.
66. Wu, J., and Woodard, R. W. (2006) New insights into the evolutionary links relating to the 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase subfamilies, *J. Biol. Chem. 281*, 4042-4048.
67. Webby, C. J., Baker, H. M., Lott, J. S., Baker, E. N., and Parker, E. J. (2005) The structure of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase from *Mycobacterium tuberculosis* reveals a common catalytic scaffold and ancestry for type I and type II enzymes, *J. Mol. Biol. 354*, 927-939.
68. Zhao, J., and Herrmann, K. M. (1992) Cloning and sequencing of a second cDNA encoding 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase from *Solanum tuberosum*, *Plant Physiology 100*, 1075-1076.
69. Dyer, W. E., Weaver, L. M., Zhao, J., Kuhn, D. N., Weller, S. C., and Herrmann, K. M. (1990) A cDNA encoding 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase from *Solanum tuberosum* L, *J. Biol. Chem. 265*, 1608-1614.
70. Gosset, G., Bonner, C. A., and Jensen, R. A. (2001) Microbial origin of plant-type 2-keto-3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthases, exemplified by the chorismate- and tryptophan-regulated

enzyme from *Xanthomonas campestris*, *J. Bacteriol. 183*, 4061-4070.
71. Silakowski, B., Kunze, B., and Muller, R. (2000) *Stigmatella aurantiaca* Sg a15 carries genes encoding type I and type II 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthases: involvement of a type II synthase in aurachin biosynthesis, *Arch. Microbiol. 173*, 403-411.
72. Guo, J., and Frost, J. W. (2002) Kanosamine biosynthesis: a likely source of the aminoshikimate pathway's nitrogen atom, *J. Am. Chem. Soc. 124*, 10642-10643.
73. Webby, C. J. (2006) Structural and funtional charaterization of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase from *Helicobacter pylori* and *Mycrobacterium tuberculosis*, Massey University, Palmerston North.
74. Tribe, D. E., Camakaris, H., and Pittard, J. (1976) Constitutive and repressivle enzymes of the common pathway of aromatic biosynthesis in *Escherichia coli* K-12: regulation of enzyme synthesis at different growth rates, *J. Bacteriol. 127*, 1085-1097.
75. Hoffmann, P. J., Catchesi.De, and Doy, C. H. (1972) Separation of three allosterically inhibitable 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthases from extracts of *Neurospora crassa* and purification of tyrosine inhibitable isozyme, *Biochim. Biophys. Acta 268*, 550-&.
76. Webby, C. J., Jiao, W. T., Hutton, R. D., Blackmore, N. J., Baker, H. M., Baker, E. N., Jameson, G. B., and Parker, E. J. (2010) Synergistic allostery, a sophisticated regulatory network for the control of aromatic amino acid biosynthesis in *Mycobacterium tuberculosis*, *J. Biol. Chem. 285*, 30567-30576.
77. Sasso, S., Okvist, M., Roderer, K., Gamper, M., Codoni, G., Krengel, U., and Kast, P. (2009) Structure and function of a complex between chorismate mutase and DAHP synthase: efficiency boost for the junior partner, *EMBO J 28*, 2128-2142.
78. Wlodawer, A., Minor, W., Dauter, Z., and Jaskolski, M. (2008) Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures, *FEBS J. 275*, 1-21.
79. DeLeo, A. B., and Sprinson, D. B. (1968) Mechanism of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate (DAHP) synthetase, *Biochem. Bioph. Res. Co. 32*, 873-877.
80. Floss, H. G., Onderka, D. K., and Carroll, M. (1972) Stereochemistry of the 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthetase reaction and the chorismate synthetase reaction. , *J. Biol. Chem. 247*, 736-744.
81. Onderka, D. K., and Floss, H. G. (1969) Steric course of the chorismate synthetase reaction and the 3-deoxy-D-*arabino*-heptulosonate 7-phosphate (DAHP) synthetase reaction., *J. Am. Chem. Soc. 91*, 5894-5896.
82. Ahn, M., Pietersma, A. L., Schofield, L. R., and Parker, E. J. (2005) Mechanistic divergence of two closely related aldol-like enzyme-catalysed reactions, *Org. Biomol. Chem. 3*, 4046-4049.
83. Glide, version 4.0, Schrodinger, LLC, New York, NY, 2005.

84. Schrodinger Suite 2006 Induced Fit Protocol; Glide version 4.0, Schrodinger, LLC, New York, NY, 2005; Prime version 1.5, Schrodinger, LLC, New York, NY, 2005.
85. Reichau, S., Jiao, W. T., Walker, S. R., Hutton, R. D., Baker, E. N., and Parker, E. J. (2011) Potent inhibitors of a Shikimate pathway enzyme from *Mycobacterium tuberculosis*: combining mechanism- and modeling-based design *J. Biol. Chem. 286*, 11.
86. Walker, S. R., Cumming, H., and Parker, E. J. (2009) Substrate and reaction intermediate mimics as inhibitors of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase, *Org. Biomol. Chem. 7*, 3031-3035.
87. Mammen, M., Choi, S. K., and Whitesides, G. M. (1998) Polyvalent interactions in biological systems: implications for design and use of multivalent ligands and inhibitors, *Angew. Chem. Int. Edit. 37*, 2755-2794.
88. Atkinson, D. E. (1966) Regulation of enzyme activity, *Annu. Rev. Biochem. 35*, 85-&.
89. Monod, J., Wyman, J., and Changeux, J. P. (1965) On the nature of allosteric transitions: a plausible model, *J. Mol. Biol. 12*, 88-&.
90. Koshland, D. E., Nemethy, G., and Filmer, D. (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits, *Biochemistry 5*, 365-385.
91. Sanwal, B. D. (1970) Allosteric controls of amphibolic pathways in bacteria, *Bacteriol. Rev. 34*, 20-&.
92. Hammes, G. G., and Wu, C. W. (1974) Kinetics of allosteric enzymes, *Annu. Rev. Biophys. Bio. 3*, 1-33.
93. Pardee, A. B., and Reddy, G. P. V. (2003) Beginnings of feedback inhibition, allostery and multi-protein complexes, *Gene 321*, 17-23.
94. Cui, Q., and Karplus, M. (2008) Allostery and cooperativity revisited, *Protein Sci. 17*, 1295-1307.
95. Kern, D., and Zuiderweg, E. R. P. (2003) The role of dynamics in allosteric regulation, *Curr. Opin. Struct. Biol. 13*, 748-757.
96. Gunasekaran, K., Ma, B., and Nussinov, R. (2004) Is allostery an intrinsic property of all dynamic proteins?, *Proteins 57*, 433-443.
97. Swain, J. F., and Gierasch, L. M. (2006) The changing landscape of protein allostery, *Curr. Opin. Struct. Biol. 16*, 102-108.
98. Bahar, I., Chennubhotla, C., and Tobi, D. (2007) Intrinsic dynamics of enzymes in the unbound state and, relation to allosteric regulation, *Curr. Opin. Struct. Biol. 17*, 633-640.
99. Goodey, N. M., and Benkovic, S. J. (2008) Allosteric regulation and catalysis emerge via a common route, *Nat. Chem. Biol. 4*, 474-482.
100. Tsai, C. J., del Sol, A., and Nussinov, R. (2008) Allostery: absence of a change in shape does not imply that allostery is not at play, *J. Mol. Biol. 378*, 1-11.
101. Tsai, C. J., Del Sol, A., and Nussinov, R. (2009) Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms, *Mol. Biosyst. 5*, 207-216.
102. Dobson, R. C. J., Griffin, M. D. W., Jameson, G. B., and Gerrard, J. A. (2005) The crystal structures of native and (*S*)-lysine-bound dihydrodipicolinate synthase from *Escherichia coli* with improved

resolution show new features of biological significance, *Acta Crystallogr. D Biol. Crystallogr. 61*, 1116-1124.
103. Das, R., Mazhab-Jafari, M. T., Chowdhury, S., SilDas, S., Selvaratnam, R., and Melacini, G. (2008) Entropy-driven cAMP-dependent allosteric control of inhibitory interactions in exchange proteins directly activated by cAMP, *J. Biol. Chem. 283*, 19691-19703.
104. Bradley, M. J., Chivers, P. T., and Baker, N. A. (2008) Molecular dynamics simulation of the *Escherichia coli* NikR protein: equilibrium conformational fluctuations reveal interdomain allosteric communication pathways, *J. Mol. Biol. 378*, 1155-1173.
105. Henzler-Wildman, K., and Kern, D. (2007) Dynamic personalities of proteins, *Nature 450*, 964-972.
106. Tuomas, V., Ferdinand, M., Chris, O., Carsten, C., and Mikael, P. (2010) Molecular mechanism of allosteric communication in the human PPARα-RXRα heterodimer, *Proteins 78*, 873-887.
107. Fatmi, M. Q., Ai, R., and Chang, C.-e. A. (2009) Synergistic regulation and ligand-induced conformational changes of tryptophan synthase, *Biochemistry 48*, 9921-9931.
108. Popovych, N., Sun, S. J., Ebright, R. H., and Kalodimos, C. G. (2006) Dynamically driven protein allostery, *Nat. Struct. Mol. Biol. 13*, 831-838.
109. Lockless, S. W., and Ranganathan, R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families, *Science 286*, 295-299.
110. Süel, G. M., Lockless, S. W., Wall, M. A., and Ranganathan, R. (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins, *Nat. Struct. Biol. 10*, 59.
111. Estabrook, R. A., Luo, J., Purdy, M. M., Sharma, V., Weakliem, P., Bruice, T. C., and Reich, N. O. (2005) Statistical colevolution analysis and molecular dynamics: identification of amino acid pairs essential for catalysis, *Proc. Natl. Acad. Sci. U. S. A. 102*, 994-999.
112. Marcelino, A. M. C., Smock, R. G., and Gierasch, L. M. (2006) Evolutionary coupling of structural and functional sequence information in the intracellular lipid-binding protein family, *Proteins 63*, 373-384.
113. Nascimento, A. S., Krauchenco, S., Golubev, A. M., Gustchina, A., Wlodawer, A., and Polikarpov, I. (2008) Statistical coupling analysis of aspartic proteinases based on crystal structures of the *Trichoderma reesei* enzyme and its complex with pepstatin A, *J. Mol. Biol. 382*, 763-778.
114. Shen, H. B., Xu, F., Hu, H. R., Wang, F. F., Wu, Q., Huang, Q., and Wang, H. H. (2008) Coevolving residues of $(\beta/\alpha)_8$-barrel proteins play roles in stabilizing active site architecture and coordinating protein dynamics, *J. Struct. Biol. 164*, 281-292.
115. Bachega, J. F. R., Navarro, M., Bleicher, L., Bortoleto-Bugs, R. K., Dive, D., Hoffmann, P., Viscogliosi, E., and Garratt, R. C. (2009) Systematic structural studies of iron superoxide dismutases from human parasites and a statistical coupling analysis of metal binding specificity, *Proteins 77*, 26-37.

116. Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009) Protein sectors: evolutionary units of three-dimensional structure, *Cell 138*, 774-786.
117. Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. Features and development of Coot, *Acta Crystallogr. Sect. D-Biol. Crystallogr. 66*, 486-501.
118. Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD: Visual molecular dynamics, *J. Mol. Graphics 14*, 33-&.
119. Kanehisa, M., and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucl. Acids Res. 28*, 27-30.
120. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs, *Nucl. Acids Res. 38*, D355-D360.
121. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG, *Nucl. Acids Res. 34*, D354-D357.
122. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007) Clustal W and clustal X version 2.0, *Bioinformatics 23*, 2947-2948.
123. Rivoire, O., and Ranganathan, R. (2008) Note 103: A summary of SCA calculations.